From the Department of Medical Epidemiology and Biostatistics
Karolinska Institutet, Stockholm 2013

Methods for deep examination of DNA

Mårten Neiman

Karolinska Institutet

Karolinska Institutet
Department of Medical Epidemiology and Biostatistics
Nobels Väg 12A
171 77 Solna, Sweden

ISBN 978-91-7549-173-8
Printed by Eprint AB 2013

Abstract

The development of sequencing technology has had a rapid pace during the last years and today, the sequencing instruments harbors enormous capacity. This thesis is about the development of methods to make the most out of this capacity and to use it for various applications.

In **paper I**, a dual tagging system for sequencing large sample sets was developed. To proof the concept, 4,700 dogs were subjected to amplicon sequencing of the 2nd exon of the gene DLA-DRB1 using a 454 genome sequencer. By using a combination of two tags, 4,992 samples can be analyzed within the same run using only 148 tagging sequences. In our experiment, 95% of the generated PCR-products achieved a sufficient read depth ($\geq 20\times$) for variant calling.

Paper II solved a problem coupled to amplicon sequencing experiments, namely contamination of sequence data from unwanted by-product formation. By hybridizing an oligo nucleotide coupled to a fluorescent dye to the wanted PCR products, fluorescence activated cell sorting (FACS) could be used to enrich the target molecules after emulsion PCR. The resulted in a nearly three-fold increase of quality reads from a sorted library compared to a non-sorted.

Since the cost of sequence data has decreased during the last years, the budget item of library preparation have became more abundant in sequencing experiments. In **paper III**, library preparation using cheap bulk enzymes was explored regarding various factors affecting process efficiency. Multiplex sequence capture was also evaluated for processing up to eight samples at the cost of one single reaction.

Using circulating tumor DNA for measuring systemic tumor load have been proposed and investigated by several studies. In **paper IV**, the utility of exome sequencing for this application was assessed. The findings suggests that the ability of detecting low-frequent alleles is insufficient for clinical use where detection levels of $< 0.5\%$ is required.

Keywords: massive parallel sequencing, cancer genomes, prostate cancer, biomarkers, library preparation, amplicon sequencing, DNA barcodes, circulating DNA, enrichment, targeted sequencing, exome sequencing.

Till Maja och Märta

List of publications

The presented thesis is based on the following four articles, referred to by their Roman numerals (I-IV). All articles are included in the Appendix.

Paper I - **Neiman M.**, Lundin S., Savolainen P. and Ahmadian A. Decoding a Substantial Set of Samples in Parallel by Massive Sequencing. *PLoS ONE* (2011) 6(3):e17785.

Paper II - Sandberg J., **Neiman M.**, Ahmadian A. and Lundeberg L. Gene-specific FACS sorting method for target selection in high-throughput amplicon sequencing. *BMC Genomics* (2010) 11:140.

Paper III - **Neiman M.**, Sundling S., Grönberg H., Hall P., Czene K., Lindberg J. and Klevebring D. Library Preparation and Multiplex Capture for Massive Parallel Sequencing Applications Made Efficient and Easy. *PLoS ONE* (2012) 7(11):e48616.

Paper IV - **Neiman M.***, Klevebring D.*, Wiklund F., Sundling S., Wiklund P., Egevad L., Grönberg H. and Lindberg J. Exome sequencing of cell-free plasma DNA in prostate cancer patients. *Manuscript*

**Both authors contributed equally to the work.*

Related publications

Lindberg J., Klevebring D., Liu W., **Neiman M.**, Xu J., Wiklund P., Wiklund F., Mills I.G., Egevad L., Grönberg H. Exome Sequencing of Prostate Cancer Supports the Hypothesis of Independent Tumour Origins. *Eur Urol* (2012) Feb;63(2):347-53.

Lindberg J., Mills I.G., Klevebring D., Liu W., **Neiman M.**, Xu J., Wikstrom P., Wiklund P., Wiklund F., Egevad L. and Grönberg H. The Mitochondrial and Autosomal Mutation Landscapes of Prostate Cancer. *Eur Urol* (2013) Apr;63(4):702-8

Contents

Preface

Developing methods for DNA research is like climbing. There is a branch within climbing called bouldering. Here, the goal is to complete a short route at a low height without the use of ropes. Since the climb is condensed, it consists of only a handful of moves, each of which is carefully defined. Therefore, within bouldering one talks about solving problems rather than completing routes. In harder boulder problems, there are often specific moves that are extra difficult and requires several rounds of trial and error to be conquered. First when the climber is capable of performing each discrete move with perfection, the problem can be completed in one sequence. The same applies when developing a laboratory procedure for solving a technical puzzle. Each step has to be carefully designed and optimized and to do this, several rounds of iteration have to be undertaken. First when every step can be performed with perfection, the protocol as a whole is ready to be applied in a biological study. Welcome to my thesis.

The field of DNA research is moving at an incredible speed and during the process of writing this book, several new findings were published and the text had actually to be rewritten. These are exciting times. Welcome to my thesis, dear reader.

Mårten Neiman
Stockholm, April 29, 2013

Scope

The papers of this thesis are about making the most out of DNA sequencing experiments and to investigate their limitations with the objective of answering biological questions. The thesis is divided into three chapters.
The first chapter is mainly about the biological matters that need to be considered when developing these kinds of methods. Understanding the underlying biology and the way of designing studies of such is necessary for identifying technological limitations and how to circumvent these.
The second chapter is about the technology of DNA sequencing. The purpose of this chapter is not to give a comprehensive description of all methods available but to discuss technologies more or less associated with the papers. Since there is much to learn from history, a historical perspective of the technical developments is also undertaken.
In the third chapter, the research behind the papers and its value to the scientific community is discussed, trying not to repeat the content of the actual papers. This book is written for a reader that is a researcher within the field of DNA sciences or other related research areas.

Chapter 1

Introduction

1.1 DNA, genes and environment

DNA has fascinated humans for decades. It was observed for the first time in 1869 [1], shown to carry the genetic information in 1944 [2] and its characteristic double helix structure was resolved in 1953 [3]. The abbreviation DNA stands for deoxyribonucleic acid; *deoxyribose* is the sugar unit that, together with phosphate, builds up the backbone, and *nucleus* means core and this is where it is found in eukaryotic cells. The DNA molecule is a polymer consisting of four subunits called nucleotides and as the letters of this text form words, the sequential order of these four nucleotides determines the genetic content. A nucleotide can be subdivided into three units: a sugar, a phosphate group and an organic base. All four nucleotides share the identical sugar unit and phosphate group but have different bases: adenine (A), guanine (G), cytosine (C) and thymine (T). The double helix structure consists of two DNA molecules twisted around each other, that are held together by hydrogen bonds between the base units of the nucleotides. A certain base can only bind to another certain base forming a base pair so that the two DNA molecules complement each other. The base A binds only to T, whereas C only to G and vice versa. One of the main features of DNA is its ability to be copied. This is carried out by an enzyme called *DNA polymerase* which uses this principle of base pairing to make a complementary strand from a single stranded DNA molecule.

The complete genetic content of an organism is called *genome* and different parts of a genome have different functions. One such function, and perhaps

the most thoroughly investigated, is the coding for proteins. In eucaryotic cells, the complete coding DNA sequence needed to create a protein is not continuously ordered, but split up into smaller coding pieces called *exons*, which are separated by noncoding pieces called *introns* and flanked by regulatory elements creating a wholeness called a *gene*. In order for a cell to make use of a gene, by producing its corresponding protein, the genetic information must first be boiled down into a single continuously coding sequence and transported to the protein producing ribosomes. This is carried out by the cell through a process called *transcription* where the genetic information is copied and spliced into a molecule called messenger RNA (mRNA), that is further processed into an amino acid chain through a process called *translation*. In addition to transcription, the genetic content of a cell can also be copied in the eve of a cell division and this process is refered to as *replication*.

This whole concept of a unidirectional informational flow, DNA->DNA, DNA->RNA and RNA->protein, is called *the central dogma of molecular biology* and was first formulated by Francis Crick in 1958. At this time, the understanding of cell mechanisms was still in its early days so what Crick stated then was basically:

> *"Once (sequential) information has passed into protein it cannot get out again."*

Since he wasn't sure whether the transfers RNA->RNA, RNA->DNA or DNA->protein existed he avoided to mention these but when the knowledge grew he was criticized for simplifying things so in 1970 he reformulated the dogma and argued that what he stated in 1958 was still valid [4].

The human genome consists of 3,137,161,264 letters (hg19) [5], or bases. If one should spell out the whole sequence using a character width of 1 mm, the human genome would span 3,137 km which is the same distance as traveling from Stockholm to Paris and back agin. In this scale, an average gene would span 27 m, the protein coding part of a gene would be 1.3 m and a new gene would occur every 300 m [6]. Like the genomes of nearly all mammals, the human genome is a so called diploid genome meaning that each somatic cell possesses two copies: one inherited from the mother and one from the father.

In February 2001, two competitive research groups simultaneously published the first draft sequence of the human genome [6, 7] revealing an unexpected low proportion of the genome being protein coding, 1.1% and 1.4% respectively. What the rest 98.6-98.1% of the genome is doing has been a subject for discussions but the fact that genes and genomes contain regulatory elements has been known since the 1950s [8, 9]. The hypothesis of *junk DNA* was introduced in 1972 by the Japanese-Korean-American geneticist Susumu Ohno [10] who, based on the weight of human and bacterial genomes, argued that a human genome cannot contain genes in a proportional amount to its size (weight) since it had been showed that a gene locus has a probability of 10^{-5} per generation to achieve a deleterious mutation and if the number of genes becomes greater than or equal to 10^5, the overall probability becomes greater than or equal to 1 which did not seem reasonable. This line of thoughts resulted in two conclusions: firstly, there must be an upper limit of how many genes a genome can harbor and secondly, the vast majority of our genome functions as a buffer for the cell to make mistakes during replication without harming its genes; the function of doing nothing.

The Encyclopedia of DNA Elements (ENCODE) project was initiated in 2004 in order to characterize all functional elements of the newly completed human genome [11, 12]. In September 2012, 30 articles related to the project were published in high-impact journals like *Nature*, *Genome Research*, and *Genome Biology*. The massive project had found that 80.4% of the genome participates in at least one biochemical event, identified 20,687 protein coding genes (1.22% of the genome) and an estimate of 60-80% of the genome was shown present in at least one transcript in at least one cell type [13]. The function of all this transcribed but non-coding DNA is either being part of the translation machinery (e.g. rRNA, tRNA), RNA processing machinery (e.g. snRNA, snoRNA) or as different regulatory RNA species (e.g. miRNA, siRNA, lncRNA). Noticeably, the vast majority of our genome is functional in some way and the hypothesis of *junk DNA* must be regarded as disproved.

Accordingly, the main function of the genome is not only coding for proteins but also to facilitate a machinery for the cell to respond to the environment and produce desired proteins in proper amounts. In other words, to regulate gene expression. It then follows that the characteristics of an organism is not only a product of its genes but also a result of the environment it lives in.

Consider two pines: one grows in a deep wood and the other in a crack of on a windblown rock close to the sea. These two trees do not shear the same appearance at all. They do share the same genome but in order to survive in their respective environment, the trees have adapted themselves. Different traits are more or less influenced from genes or environment. For example, if an organism is struck by the lightning and injured, its genes does not have any influence in this (unless it is a giraffe) but the injure is a pure effect of the environment. The sex, on the other hand, is completely determined by the genes in (almost) all mammals. But most traits need both genetic and environmental factors to occur. For example obesity, which prevalence is not only affected by life style factors, such as lack of physical exercise or abundance of food but there is also an inheritable genetic component [14].

1.2 Genetic variation

The ability to vary genetic content is the mechanism enabling the never ending iterative process called evolution. The offspring of organisms is not genetically identical to its ancestors and some of its abilities might be sophisticated, extended or lost. The environment decides if the offspring's genetic set-up was beneficial or not. If abilities like feeding, escaping predators or breeding were improved, these variants are likely to be incorporated in the growing population whiles harmful variants are less likely to survive.

When a cell divides, the genome is replicated. The replication machinery is not perfect and errors occur during this process. If the new cell is a germ cell, these errors will be passed on to all cells of the offspring and through this, genetic variation is introduced. In 2007, the project *1000 genomes* was initiated with the aim of making a comprehensive map of genetic variation in humans. In the pilot phase of the project, two trios of mother-father-child were sequenced at high coverage using multiple platforms in order to estimate the germline single nucleotide mutation rate. After confirmative resequencing experiments they found 35 and 49 germline mutations respectively, which yielded an estimated mutation rate of 10^{-8} per generation and base pair [15].

The word mutation comes from the latin word *mutatio* which means *change* and the most abundant type of genetic variation is point mutations followed by short insertions and deletions (indels) and structural variation. When

studying individual genomes it is preferable to compare them to a reference genome, which represents a consensus of the most common variants of all positions, and look at the deviations from this. The term SNP (single nucleotide polymorphism, pronounced *snip*) was first defined as common nucleotide substitutions with a minor allele frequency (MAF) of $\geq 5\%$ but nowadays, one also talks about *rare* SNPs and the term could basically be interpreted as singe nucleotide variation (SNV) of any frequency.

If the genome of a randomly chosen (healthy) individual was compared to the reference sequence one would find around 3×10^6 SNV's out of which far from all are expected to be functional. To clearify the picture, the regional positions of the SNV's can be used to estimate their functional impact. When comparing genomes from different species, some regions (not only within genes but also intergenic) are said to be conserved, which means that they are highly similar even between distantly related organisms and since they have been kept intact throughout evolution they are thought to be functionally important. To estimate the fraction of functional variants, the 1000 genomes project investigated how many SNV's were situated at conserved positions. To further evaluate the SNV's damaging potential they looked in *The Human Gene Mutation Database* [16] and at how common the SNP's were (common variants are expected to be less harmful because of evolutionary pressure). A SNV's prevalence can be subdivided in common variants ($> 5\%$), low-frequent variants ($0.5 - 5\%$) and rare variants ($\leq 0.5\%$). Strikingly, it was shown that healthy individuals carry 2-5 potentially damaging rare variants and 1-2 rare variants previously associated with some form of cancer (see table 1.1) [17].

Variant type	**Common**	**Rare**
Non-synonymous	2,500	130-400
Damaging	20-40	2-5
Loss-of-function	150	10-20
Cancer associated	-	1-2

Table 1.1 Expected single nucleotide variants in conserved regions for a randomly selected healthy individual estimated by the 1000 genomes project.

So, how can people still remain healthy? There are two copies of each gene meaning that there is a backup copy. Also, several genes can have similar functions, so called genetic redundancy. However as mentioned earlier, not only our genes determine our health but also our environment, life-style and stochastic factors (also known as luck).

1.3 Studying genes and traits

Since long before the genetic information carrying role of DNA was discovered, efforts have been made to associate inherited genetic variants with physiological features. The earliest associations were made long before any technology for genotyping was available and instead, so called linkage analysis was used. *Linkage* is the phenomenon of genetic components being inherited together because of their physical nearness. Among the first observations were the sex-linked traits of red-green color blindness and haemophilia [18].

There are different genetic mechanisms behind different traits and a common way of categorizing is by subdividing them into single gene causing (mendelian) traits and complex traits involving several genes. Completely mendelian phenotypic traits are rare but some examples are dry earwax (first visible trait associated with a SNP) [19] and a form of albinism [20]. A lot of traits have been thought to be mendelian but have later been shown to be complex. An example is eye color were the gene OCA2 accounts for 74% of human eye color variation [21].

Finding the gene behind a mendelian trait is easier than trying to understand the complete mechanism behind a complex trait, as one single gene has a much larger effect size than each of the components of a complex trait. *Effect size* is the amount of impact contributed by a gene on a trait. The smaller the effect size, the larger sample size is needed to gain sufficient statistical power to detect it (see figure 1.1).

The size of the human genome, its many polymorphic positions it contains and the fact that it is very expensive to analyze the whole genome in thousands of samples makes it challenging to associate genetic variants and traits. Since investigating the whole genome is unfeasible, the search has to be restricted to selected parts rising the question: where to look? Currently,

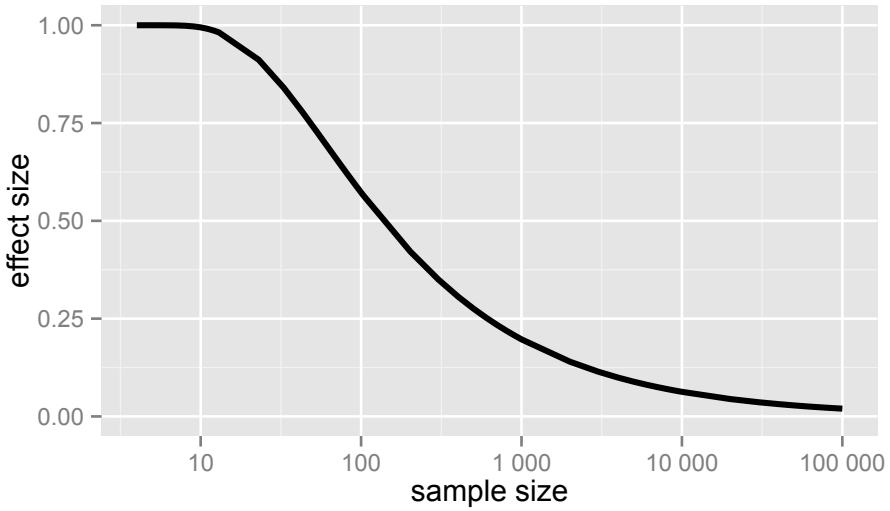

Figure 1.1 Correlation between sample size and the smallest detectable effect size at a power of 0.8 and a significance level of 5×10^{-8} given equal amounts of cases and non-cases.

there are two approaches for answering this question. Firstly, genome-wide features like SNP's or coding regions can be targeted. This is the more general approach since it does not require any previous knowledge about the biological functions behind the trait to be studied but it is restricted to previously known SNP's or genetic elements of similar function. Secondly, genes previously associated with the trait of interest, or genes having a biochemical function possibly related to the trait can be pinpointed. This second approach, commonly referred to as *fine mapping*, requires more knowledge in order to select interesting regions but allows for finding rare variants and investigating many types of distinct regions. Nevertheless, a genome-wide study could be followed by a fine mapping study to refine the first's findings.

The outcome from a genetic association is an estimate of the relative probability that a person with a certain gene variant develops a certain physiological feature, like a disease, compared to a person without the variant. This information could be used to identify individuals with high risk and change their life style accordingly and perhaps doing health checkups more often. The information can also be of use in screening programs where persons at high risk can be screened more often than people in general and if protecting

markers are known, these low-risk individuals can be assessed more seldom. Another important use is when the dose of medical drugs is determined since both the rate of degradation and the degree of side effects are associated with genetic variants [22–24].

During the first years of the new millennium, high density SNP arrays entered the market making genome wide SNP scans possible [25–27]. The earliest versions were able to detect allelic imbalances in 600 positions [25] and since then, the density of the arrays have grown and today (early 2013), Illumina offers a chip capable of analyzing up to 4.8 million features, corresponding to one marker per 680 bases throughout the genome, in four samples simultaneously [28]. These platforms have been the basis of genome-wide association studies (GWAS) where often thousands of individuals are genotyped in millions of genetic loci.

Because of the fact that only previously known SNP's, common and low-frequent, are included in a GWAS (since they have to be known when the SNP-chip is designed) and that these levels of occurrence are seldom alone coupled to disease states, highly penetrant variants are rarely found (see figure 1.2). This is a consequence of the *common disease/common variant (CD/CV) hypothesis* discussed by David Reich and Erik Lander in 2001 [36]. It states that common diseases are caused by common variants and their logical reasoning is that the effect size of each common variant for a common disease must be small relative to those of rare disorders.

Because of the very high number of markers that are tested for association in a GWAS, some precautions regarding significance level must be taken. Traditionally, a significance level of $alpha = 0.05$ is regarded as sufficient, meaning that the *null* hypothesis of no correlation is rejected if the p-value is less than 0.05. This means that in 5% of the tests, the *null* hypothesis is rejected even though it is true. In a GWAS where one million SNP's are tested, each with a 5% probability of its *null* hypothesis being falsely rejected, the false discovery rate would be inadvisable high. One way of handling this is by requiring a higher significance level. When choosing an appropriate level for a specific study, factors like number of included SNP's and the distribution of linkage disequilibrium within the studied population are to be considered. As an example, the genome-wide two-sided significance threshold for an Eu-

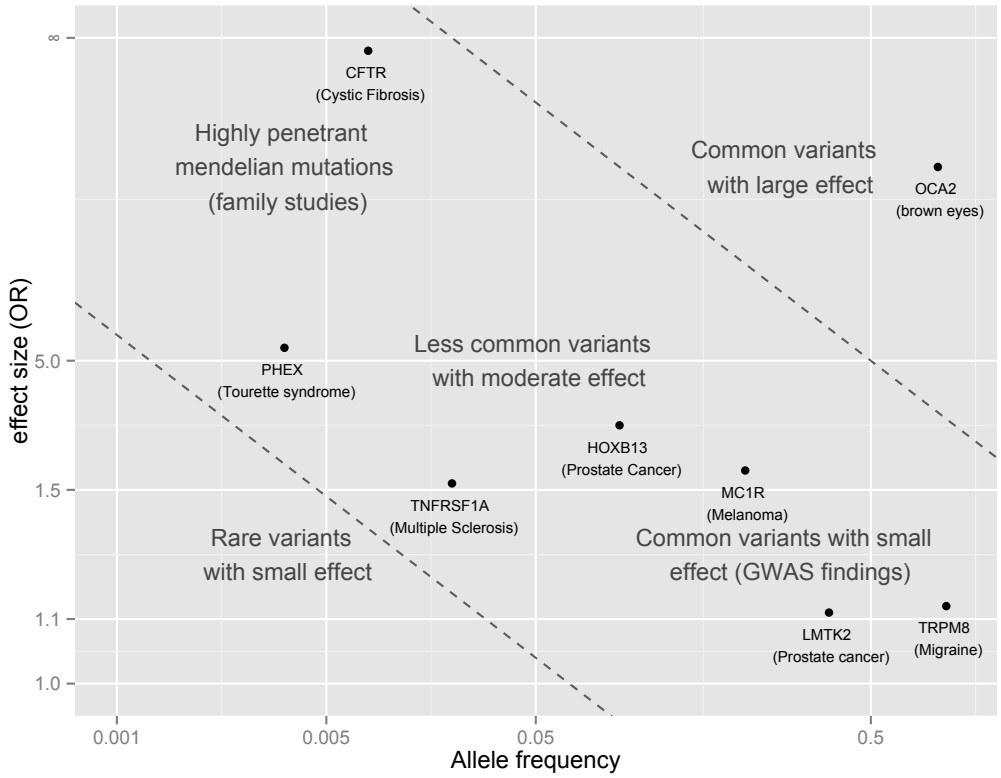

Figure 1.2 The effect size of associated genes in relation to the allele frequency in control samples [29–34]. A major proprtion of genetic findings tend to lie between the dashed lines [35].

ropean Caucasian population has been estimated to 7.2×10^{-8} [37].

Another computational challenge in a GWAS is performing *multi-locus analysis*, that is analyzing interaction effects between SNP's. Testing pair-wise interaction among one million SNP's demands a lot of computational power so some kind of filtering is preferred. A common strategy is to first perform a single-locus analysis and then move on into multi-locus analysis using the most significant findings. However using this strategy, there is a risk of missing significant interaction between SNP's that does not influence risk of disease alone [35].

The GWAS strategy have been successful, enabling the association of thousands of genes and traits. As of April 25, 2013, there was 1,577 publications

and 9,914 SNPs in the Catalog of Published Genome-Wide Association Studies supplied by The National Human Genome Research Institute [38].

1.4 Cancer genomes

Cancer is a genetically driven disease. It occurs through cumulative somatic mutations in tissues. The word *somatic* is derived from Greek and means *of the body*. Somatic mutations occur after conception, meaning that they are not shared between all cells of a body but only within a subpopulation of cells that have a common ancestor where the mutation occurred. In contrast, germline variation occurs at meiosis and is inherited to the offspring which carries the variation in all of its cells. If the variation is not damaging in a way that makes mating non-feasible, it can be incorporated to a growing population (discussed in chapter 1.2).

Cancer occurs when the accumulated somatic mutations of a cell causes an increased division rate and over time, a tumor is grown. These mutations occur when cells divide in our self repairing and renewing bodies because of limitations of our DNA replicating machinery. The number of cell divisions that takes place in the human body during a normal life time has been estimated to 10^{16} [39], which gives a high probability of some cell collecting sufficient advantageous mutations for developing a tumor. Mutations can also occur from the external exposure of DNA damaging agents, such as ionizing radiation, ultra violet (UV) radiation, tobacco smoke and chemicals. The number of mutations carried by a cancer tumor varies between different types. It has been estimated that a typical solid tumor (such as one derived from the colon, breast or pancreas) has 33 to 66 genes mutated so that their coded amino acid sequence is altered (non-synonymous mutations) [40]. Tumors located in tissues more exposed to DNA damaging agents, such as skin and lung tissue, are also more extensively mutated. For example, a lung cancer tumor in a smoking patient harbors ten times as many mutations as the tumor of a non-smoking patient [41]. The number of somatic mutations in the cells of the body are accumulated over time and luckily, it takes a while for a cancer cell to evolve. This is why the risk of cancer increases with age.

Naturally, not all accumulated somatic mutations of a tumor are causative. When studying somatic variation in cancer, it is common to categorize them

into *driver mutations*, mutations that contributes to the cause, and *passenger mutations*, mutations also carried by the tumor but without affecting its properties. A lot of effort has been made to map and understand which genes are critical for tumor development and so far, approximately 140 genes have been identified as *driver* genes, which means that they can drive tumorigenesis if they are mutated. These 140 genes are involved in three main functions of the cell: genome maintenance, cell survival and cell fate and an average tumor has two to eight of these genes mutated [40]. Cancer driver genes are classified as *proto-oncogenes* or *tumor suppressor genes*. If a proto-oncogene is genetically altered, it can be activated into an *oncogene*. The activation makes the gene over expressed or hyperactive in some way, either by chromosomal rearrangements, where the proto-oncogene is combined with a more active transcription start site and/or enhancer elements, by point mutations in regulatory/coding elements making the protein product more abundant/active or through gene amplification [42]. The proteins of oncogenes are often part of the proliferation control machinery or apoptosis regulation and are the targets of several cancer therapies.

The analysis of cancer genomes may have two aims: studying one patient or studying the disease. The goal of investigating the cancer genome of one patient could be to characterize tumor specific genetic alterations in order to tailor therapy and/or biomarkers. To do this, DNA from tumor tissue is investigated and compared to germline (normal) DNA, often derived from blood, so that only somatic tumor specific variation is considered in the analysis. In order to study the disease, a cohort of patients needs to be investigated, often as tumor-normal pairs, with the goal of identifying biomarkers, characterize clinically relevant subtypes, identify therapeutic drug targets or to simply achieve a deeper understanding of the disease mechanisms [43].

1.5 Biomarkers

According to the National Cancer Institute (NCI) Dictionary of Cancer Terms [44], the definition of a biomarker reads as:

> *"A biological molecule found in blood, other body fluids, or tissues that is a sign of a normal or abnormal process, or of a condition or disease. ..."*

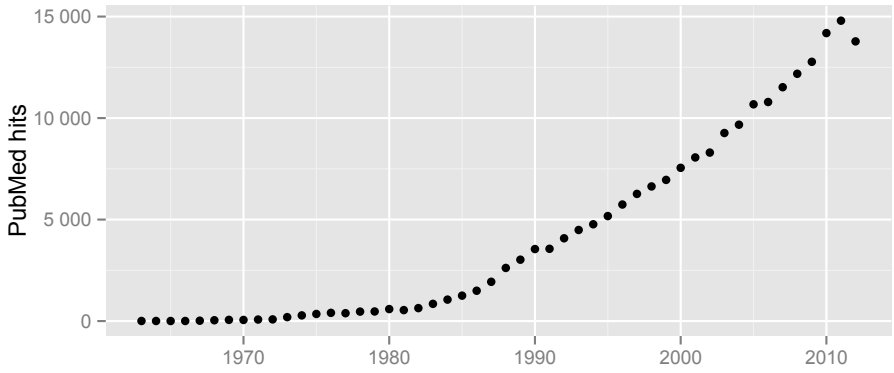

Figure 1.3 Number of hits per year in PubMed from the search string: *"cancer biomarker"*.

This molecule might be a protein, DNA, RNA or a small metabolic product. Today, several biomarkers for various forms of cancer have been characterized [45]. Generally, a biomarker can provide several sorts of information:

- **Risk** - a marker can tell if a patient has an increased (or decreased) risk of developing disease. This is often a genetic variant but can also be a physiological trait like mammographic density [46].
- **Diagnosis** - the marker tells if the patient suffers from disease or not. A blood based marker has the potential of detecting a tumor before it becomes palpable.
- **Prognosis** - disease is a fact but how aggressive is it? Continuously measurements of a biomarker could monitor tumor burden, therapy response and prevent over-treatment.
- **Relapse** - when the patient is cured a biomarker could tell if the disease returns.

As shown in figure 1.3, the number of publications with the keyword "cancer biomarker" has grown for each year and in 2011, more than 40 new papers were published per day. But only a handful of these (less than 3 per year between 1994-2003 [45]) are approved by The US Food and Drug Administration and implemented into clinical practice. There are two main categories of biomarkers that fails to reach the clinic: firstly, those that are true discoveries but with poor performance so that their contribution of information is of

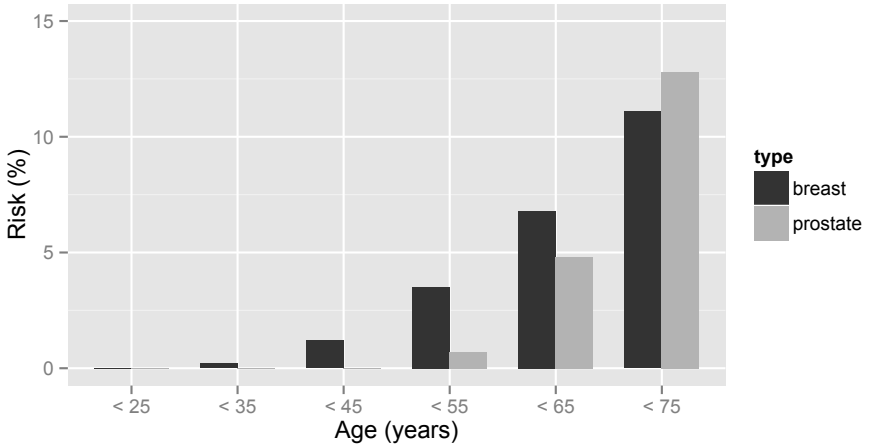

Figure 1.4 Cumulative probability of developing breast or prostate cancer before a certain age among Swedish women and men respectively showing that the risk of prostate cancer comes later in life. Data from [49].

limited clinical use and secondly, those that are false discoveries (even thou they are statistically significant findings). To aid the assessment of quality of a tumor marker study, guidelines for providing relevant information and using appropriate scientific methods have been proposed [47]. The goal of this was to promote complete reporting and make it easier for other scientist to understand the context of the results. The largest hurdle for new biomarkers to reach the market is that, for being worth the investment to take them into the clinic, they have to perform significantly better regarding sensitivity and specificity than existing biomarkers. This means that they have to provide such vital information that decisions of therapy can be made more accurately and so that the clinical cost of doing the actual test can be motivated by savings in over/under treatment [48].

1.6 Prostate cancer

In Sweden, prostate cancer is the most common form of cancer followed by breast cancer. In 2011, 9,663 patients were diagnosed with prostate cancer which corresponds to 34.2% of all cancer cases in men [49]. The relative 5-year survival was 87.3% and 10-year survival was 68.5%. The disease occurs

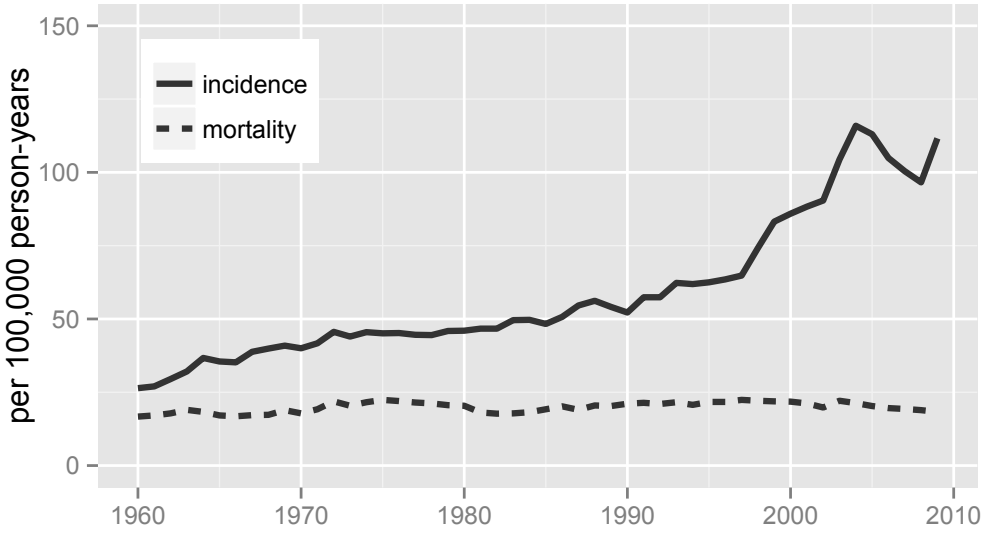

Figure 1.5 The yearly incidence and mortality of prostate cancer in Sweden per 100,000 person-years standardized to the World Standard Population. Data from NORDCAN [51].

relatively late in life, even for being a cancer and compared to breast cancer, the cumulative risk of being sick before a certain age becomes notable about ten years later than in breast cancer (see figure 1.4). Only 85 cases where the patient's age at diagnosis was ≤ 50 years were reported during 2011 in Sweden and the mean age at diagnosis is 72-74 years [50]. The annual incidence of prostate cancer in Sweden has doubled since the 1970's from a little less than 50 cases per 100,000 person-years to around 100 cases per 100,000 person-years (age standardized to the World Standard Population) and about 2.1% increase per year over the last 20 years (see figure 1.5) [49, 51]. This increase is much due to the introduction of the PSA test (discussed later in this chapter) in the 1990's but during the last five years, the incidence of prostate cancer has somewhat stabilized whiles the mortality has remained relatively unchanged since the 1960's [49].

The most established risk factors for having prostate cancer are ethnic origin, age, family history and high concentrations of insulin growth factor I (IGF-I) in blood. The three former factors cannot be affected through

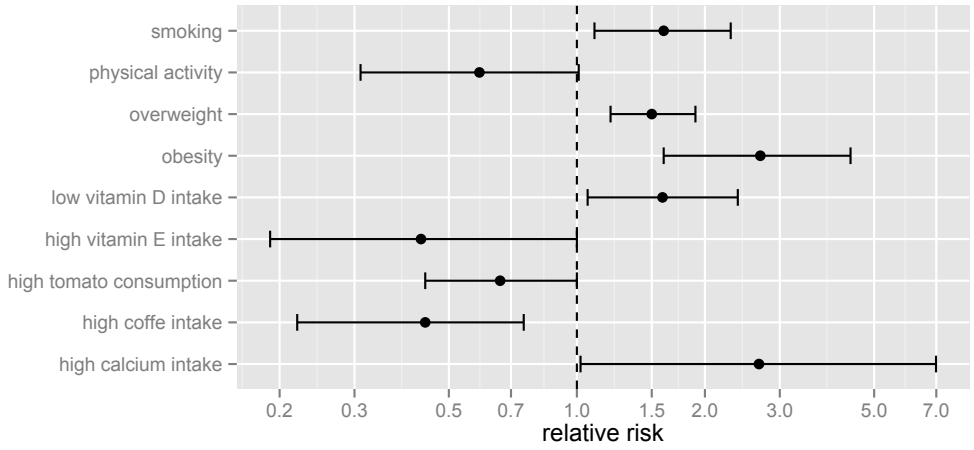

Figure 1.6 The relative risk of various life-style and dietary factors investigated for **lethal prostate cancer**. Since these estimates comes from different studies with different design and definitions of outcome, between variable comparisons should be done with caution. Data obtained from [52].

life-style, while the latter is thought to be affected by fat consumption since a raised insulin production also results in higher IGF-I levels [50]. Several life-style and dietary factors have been investigated for affecting the risk of having prostate cancer but so far, mostly conflicting or negative correlations have been observed. Fat intake has been shown to have a minor effect (OR: 1.4, 95%CI: 1.1-1.8) [53] as well as calcium (RR: 1.34, 95%CI: 1.04-1.71) [54]. Nevertheless, if the outcome is defined as *lethal* prostate cancer instead, several factors have been shown to have an effect. This could be because they have influence on the effectiveness of therapy or that several biological pathways are activated in the progression to lethal disease and hence, more opportunities for interaction [52]. In figure 1.6, the effect from some environmental factors on the risk of having lethal prostate cancer have been plotted.

As mentioned, family history is an established risk factor for prostate cancer. Having a first-degree relative diagnosed with prostate cancer means a doubled risk (RR: 2.5, 95%CI: 2.2-2.8) and having two first-degree relatives, the risk is almost fourfold (RR: 3.5 95%CI: 2.6?4.8) as estimated through a meta-analysis done in 2003 [55]. As seen, the risk is affected by the number

of affected relatives but it is also affected by the age at diagnostics of the affected relative [56]. These strict correlations suggests a high heritability of prostate cancer and a twin study has estimated that more than 40% of its susceptibility is due to genetic factors [57]. In spite of this, no highly penetrant genes have been identified and the inheritance pattern of prostate cancer has been found to be more genetically complex. Several GWAS have successfully attempted to find association between common SNP's and prostate cancer incidence and until recent, almost 50 loci had been identified with low to moderate effect (OR between 0.83-1.79) [58]. In April 2013, a large international consortium called PRACTICAL presented their findings from genotyping approximately 25,000 cases and controls using a custom array (iCOGS). They had found 23 new loci associated with prostate cancer and more than 70 loci in total [59]. One of the most penetrant loci found so far is a rare (MAF: 0.1% in the U.S.) single nucleotide variant located in the gene HOXB13 identified by exon sequencing of 202 susceptibility genes in an american study (OR: 9.5, 95%CI: 3.5-803.3) [60]. This findings have been verified by a Swedish study (OR: 3.5, 95%CI: 2.4-5.2) [30] and another american study (OR: 4.42, 95%CI: 2.56-7.64) [61].

Prostate cancer is more or less symptomless at an early stage since the tumor is often located in the peripheral zone of the prostate [62]. The prostate gland is located just below the urinary bladder and surrounds the urethra. In a healthy person, the prostate has the size of a walnut and a mass of 11 g [63]. It functions as a secretor of slightly alkaline fluid that is a major part (50-75% of volume) of semen [64]. Because of the inaccessible placement of the prostate, a tumor is not palpable in the same degree as for example a breast tumor. At a later stage, large tumors can affect the urethra making urination difficult for the patient. However, this symptom is also coupled to benign prostatic hyperplasia, a very common, non-malignant disorder among older men [65]. It is not uncommon that a person lives his life with prostate cancer without clinical symptoms and eventually dies from other causes. In a recent study, the prostates of 340 deceased trauma patients were examined (age 1-81 years) and 41 prostate cancer cases were found. The prevalences observed for each age strata were $\leq$ 39 years: 0.7%, 40-49 years: 0%, 50-59 years: 23.4%, 60-69 years: 34.7% and 70-81 years: 45.5% [66]. When severe symptoms of prostate cancer occurs, it is often at a late stage of aggressive disease when bone metastasis causes pain.

Because of the lack of symptoms at an early stage, the need for biomarkers or early imaging of prostate cancer is imminent [67, 68]. One of the most famous markers for prostate cancer is a protein called prostate specific antigen (PSA) and its prognostic value was established for the first time in 1985 [69]. PSA is a major protein in semen and is produced by secretory epithelial cells of the prostate. The expression of PSA is positively regulated by the androgen receptor meaning that it is responsive to steroid hormone levels [70]. However, a raised PSA-level does not necessary have to be due to a tumor but is also associated with benign prostatic hyperplasia [68, 71]. This means that a large proportion of those without prostate cancer can still have high PSA levels.

There are PSA-based screening programs for enhancing detection of prostate cancer at an early stage. There has been (and still is) a massive debate on the use of these programs since the specificity of PSA is rather low, causing healthy persons being needlessly investigated for prostate cancer and over diagnosis of non-lethal cancer [72, 73]. According to an estimation of the benefit from the PSA screening between 1986-2005 in the U.S., more then 20 persons had to be diagnosed with prostate cancer for every man who actually benefitted from it [73]. Additionally, a randomized trial study conducted in Sweden concluded that for preventing one prostate cancer related death, 293 men needs to be invited for screening out of which 12 have to be diagnosed [74]. Nevertheless, an European randomized study, including 162,388 participants, showed a 20% mortality reduction in the screening group during 9 years of follow-up [67]. Also, as a diagnostic test and as a marker for disease progression, PSA has proved to be very useful.

Initially, when a patient is investigated for prostate cancer, a PSA-test is taken, the prostate is examined by a physician through rectum and sometimes, an ultrasound examination is done. If the suspicion remains after these proceedings, biopsies of the prostate are taken and examined by a pathologists who assigns a so called Gleason Score based on the microscopical evaluation of the prostate glands [62, 75].

Tumors restricted to the prostate gland, so called localized disease, are divided into three risk groups of high, medium and low risk based on PSA-value,

tumor size and Gleason Score [76, 77]. Patients that have localized prostate cancer are treated with radiotherapy, radical surgery or active surveillance depending on risk group, age and general condition. Tumors spread to adjacent tissue or metastasized to other organs, such as lymph nodes or bone, are primary treated with hormonal therapy [78].

1.7 Circulating tumor DNA

It has been known for decades that cancer tumors release cells in the blood stream that are transported to other organs and seeding new tumors, so called metastasis. This occurs at a relatively late stage of the disease but the tumor leaves molecular traces in the blood also at earlier stages. In general, tumor cells are more biochemically active than normal cells and as a consequence, they also release a larger quantity of matter to their surroundings. This might be DNA, RNA, proteins and small metabolic molecules and if they can be distinguished from molecules released from normal cells or are present in abnormal quantities, they might provide information regarding tumor characteristics. While proteins and metabolic molecules are mainly investigated in a quantitative manner, nucleic acids can carry tumor specific features within the molecule that can be confidently detected using modern technology.

The presence of free circulating nucleic acids in blood has been known since 1948 [79] and they are thought to originate from macrophages releasing DNA after engulfing cellular debris from necrotic cells [80, 81]. But it was not until 1994 that its importance as a diagnostic and prognostic cancer marker was revealed when tumor specific mutations in the oncogenes *KRAS* and *NRAS* were detected for the first time [82, 83]. It had already been observed that the total concentration of circulating DNA in blood from cancer patients were significant higher than in healthy donors, but the variation within the groups was large [84]. The concentration range of circulating DNA in blood have been estimated to 0-100 ng/ml among healthy individuals with a mean of 30 ng/ml and between 0 and >1000 ng/ml among cancer patients with a mean of 180 ng/ml [85]. The findings from studies attempted to investigate the amount of DNA in blood have been very inconsequent and this uniformity is thought to be related to the different methods used for DNA quantification, but even a large prospective study observed untenable large

variation proposing that it was due to variation in sample treatment and/or variation between the populations studied (participants recruited from all over Europe) [86].

However, it has also been shown that high levels of circulating DNA in blood does not necessary reflect malignancy. It can also be the product of any pathological process involving apoptotic and necrotic cells, such as inflammatory diseases and tissue trauma [85]. Another source of circulating DNA is the fetus of pregnant women [87, 88]. By studying the presence of Y-chromosomal DNA in the blood stream of women carrying a male fetus, it is easy to distinguish fetus DNA from mother DNA and to study the dynamics of circulating DNA. Sequential sampling of eight women that had delivered male babies showed a rapid clearance of fetal DNA with an estimated mean half-life of 16.3 min (range: 4-30 min) [89]. These studies were made during the end of the 1990's and since then, the technology has been more sophisticated and recently, a paper was published where whole-genome sequencing of mother, father and the circulating DNA from the mother had been carried out, making it possible to reconstruct the genome of the unborn fetus [90].

The degradation process of circulating DNA is not yet fully understood but experiments have shown that different forms of DNA survives for different amounts of time. By injecting purified DNA into mice, a longer survival of double stranded DNA (dsDNA) than single stranded DNA (ssDNA) was observed. It was also demonstrated that a closed ring survived longer than linear DNA [91]. Nucleases active in the blood stream have been proposed to be the major source of degradation together with the liver and macrophages as well as the kidneys since foreign DNA has been observed in the urine of both pregnant women and cancer patients [85].

Quantifying the total amount of circulating DNA present in the blood of a cancer patient cannot provide evidence of disease progression nor prognosis alone because of the large variation among individuals. Nevertheless, it has been shown that the total level of circulating DNA decreases after successful radical surgery and from response to chemotherapy [92]. The key to use blood as a liquid biopsy is to be able to discriminate between tumor DNA and normal DNA in some way.

As mentioned in chapter 1.2, the most common form of genetic variation is single nucleotide substitutions. This is also valid for tumors so looking for tumor specific SNV's in the circulating DNA content is a commonly used way of detecting tumor derived DNA and this was also the first characteristic used [82, 83]. KRAS is the most commonly assessed gene for detection of tumor specific SNV's in circulating DNA studies followed by TP53 (see table 1.2). Other types of alterations used are structural rearrangements, CNV's, microsatellites and methylation. One of the greater challenges when

Gene	Publications
KRAS	36
TP53	28
APC	4
NRAS	1

Table 1.2 Number of published studies up until 2007 where the respective gene was assessed for tumor DNA detection in plasma/serum and other body fluids using SNV's. Data obtained from [85].

detecting circulating tumor DNA is the relatively high levels of background DNA derived from normal cells and this puts a high demand on the technical approaches used. Also, different types of genetic variation provides different conditions for detection and have also various potential of reaching satisfying sensitivity and specificity.

Assessing tumor dynamics by monitoring tumor DNA in blood samples has great potential for improving clinical practice. If a genetic profile of a tumor is made, an assay specific for that particular tumor can be designed and applied for monitoring of the disease [80, 81, 93, 94]. This could enable tracking of a patient's response to various treatments, early detection of relapse and more reliable prognosis. Having these possibilities, choice of treatment and dose can be individually adapted avoiding over treatment and unnecessary side effects. For instance, radio therapy of breast cancer patient has been correlated to increased risk of ischemic heart disease [95]. Circulating tumor DNA has also been used for detecting acquired resistance to treatment by screening for KRAS mutations in colorectal cancer patients subjected to anti-EFGR therapy [96, 97]. Its utility as a biomarker for monitoring treatment

response within metastatic breast cancer patients has been benchmarked against other biomarkers such as cancer antigen 15-3 (CA 15-3) and circulating tumor cells showing that it is both more sensitive and more informative than the others [98].

Another challenge for the utility of circulating tumor DNA is the heterogeneous nature of primary tumors. Most solid tumors consists of a multitude of genetically different sub-clones [40] all of which might differ in their abilities to shed nucleic acids to their surroundings [85]. The tumor DNA measured in the blood is a mixture of the emissions from all tumor cells and its composition reflects both mutation prevalence within the tumor and the DNA secreting activity of the respective sub-clonal tumor cells. When assessing a tumor's genetic profile by sequencing a biopsy specimen, there is a risk that the biopsy missed the part of the tumor that contributes the most to the circulating DNA content and sequentially, when using this genetic profile for monitoring tumor dynamics the outcome might be erroneous.

Circulating tumor DNA can also be used for screening of low-frequent mutations within known oncogenes, enabling detection of tumors at an extremely early stage. In contrast to detection of pre-determined mutations, the positions of the mutations in a screening situation are unknown and the statistical significance level has to be adjusted accordingly making it harder to detect ultra-low frequent variants. In a study using targeted ultra-deep sequencing of plasma samples investigating a region of 5,995 bases, allele frequencies of 2% were confidently detected and mutations previously missed by biopsy sequencing could be identified [99].

Chapter 2

Technology

For every leap in technology, a leap in biological knowledge follows. This is only natural since understanding builds on making observations and drawing conclusions. If a new technology enables previous unable observations, these will be followed by previously undrawn conclusions.

2.1 Amplification

The ability of making copies of DNA molecules has been the key to modern DNA science. This process is commonly referred to as *amplification* and there are several methods for achieving this, both *in vivo* and *in vitro*, all with different abilities, implications and limitations. Not only enabling a stronger signal when interrogating DNA in different ways but amplification is also a way of lowering the complexity by targeting and only copying the region of interest.

Cloning

The first step towards the ability of copying DNA was taken in 1972 when antibiotic resistance was transferred to a strain of *E. coli* [100]. This was the beginning of molecular cloning and the following year, the use of a plasmid for carrying the genetic material into the hosting organism was demonstrated [101]. The term *clone* is said to be derived from the Ancient Greek and means "twig" (a thin terminal branch of a tree) and it was introduced in analogue to the phenomenon of growing a new tree by planting a twig. By transfecting cells with a piece of DNA coupled to a gene vital for surviving in the growth

environment, such as antibiotic resistance, this piece is copied as the cells multiplies and will be present in all resulting cells. This amplification is *clonal* meaning that within each resulting colony, all cells will be carrying the same piece of DNA received by their common ancestor by transfection. Until the mid-80's, this was the main amplification method used by the scientific community and it is a tedious and time-consuming process.

Polymerase chain reaction

In the discussion section of a paper about the replication properties of DNA polymerases, the norwegian scientist Kjell Kleppe described how a cyclic thermal reaction and a two-primer system could replicate a specific strand of DNA [102]. This was in 1970 and probably because of the lack of an effective way of synthesizing oligonucleotides in a sufficient scale to be used as primers at that time, he was never able to test his theory in real life.

Instead, it was a californian biochemist named Kary Mullis that experimentally proved the method in 1987 and he named it polymerase chain reaction (PCR) [103]. He was awarded the 1993 Nobel prize in chemistry for this achievement. In the article describing PCR, the mesophilic (likes moderate temperatures) enzyme Klenow fragment of *E. coli* DNA polymerase I was used and since it was destroyed in the heat of the denaturing step, new enzyme had to be added for each cycle. This made the process very time consuming and expensive and it was not until the thermophilic enzyme Taq DNA polymerase was brought into use that the full potential of the method was released [104].

PCR is a thermal cyclic reaction where each cycle consists of three steps conducted at three different temperatures and theoretically, the number of DNA molecules is doubled for each cycle resulting in an exponential amplification. The reaction is usually cycled 15-30 times giving 2^{15} - 2^{30} = 32,768 - 1,073,741,824 copies.

The specificity of PCR is established by the requirement of the primer-pair being complementary to the flanking regions of the amplification target. If one primer miss-anneals to another genomic region, this will result in a linear amplification generating a low-yield product and the risk of this happens is

relative high. Nevertheless, the risk of having miss-annealing to two regions close enough to each other to facilitate exponential amplification generating a high-yield bi-product is relatively low [104]. This risk is increased if several primer-pairs are present in the reaction for a multiplex amplification which is also the risk of primer-dimer formation. The latter is when two primers partly anneals to each other and creates a short highly effectively amplified by-product which tends to steal chemical capacity from the main reaction.

In summary, PCR has fundamentally revolutionized the filed of molecular biology. Today, there is not a single biotechnological facility without a PCR-machine.

Multiplex amplification

By adding several primer pairs to the PCR, a multitude of genomic regions can be amplified simultaneously saving time, DNA sample and reagents. The primer design for a **multiplex PCR** however requires significantly more attention relative a simplex reaction. All pairs must be suitable for one common annealing temperature and if all the products should be distinguishable in an electrophoresis analysis, it is preferable that they all have different sizes. As mentioned above, the probability of having by-products increases with the degree of multiplexity which further complicates the primer design. Another phenomenon seen in multiplex PCR is recombination between the different amplicons resulting in chimeric products [105]. Multiplex PCR was first described in 1988 in a setting for detecting deletions in the dystrophin gene causing the severe muscle degrading disorder Duchenne muscular dystrophy [106].

One of the most famous applications of multiplex PCR is DNA profiling for forensic and relationship testing. These tests rely on a genetic feature called short tandem repeats (STR) which are short repetitive sequence elements of 3-7 nucleotides. STR's are scattered all over the genome and because the number of repeats differs between individuals, these are a rich source of highly polymorphic markers distinguishable by size separation [107–109]. Today, the Swedish National Laboratory of Forensic Science (SKL) uses multiplex PCR for amplifying 16 different regions (15 STR's plus Amelogenin for sex determination) and four-color fluorescent aided detection by capillary

electrophoresis. The risk of two siblings having matching profiles using this kit is 1 in 50,000 [110].

It is hard to design a multiplex PCR for more than about 20 amplicons due to by-product formation and primer-dimer amplification. For amplification of a greater multitude of targets, like hundreds or even thousands, some kind of mechanism has to be used for repressing the formation of unwanted products and favoring amplification of targets. Several methods for highly multiplex amplification have been developed and there is always a tradeoff between the amount of unspecific products formed and the fraction of targets that are amplified.

Molecular inversion probes (MIP) uses a circularization procedure of the target molecules by hybridizing target specific ends of a probe on each side and then joining the ends by fill-in followed by ligation (early versions only relied on ligation). Remaining linear DNA is degraded and the circular probes are opened and amplified using universal primers [111]. MIP's have been implemented in a variety of applications such as SNP genotyping [112, 113], CNV profiling [114] and exome enrichment (55,000 targets in one single reaction) [115, 116].

The company Illumina have developed a method named **Golden Gate** for multiplex amplification and genotyping. At genomic level, an allele specific extension from probes carrying allele specific amplification handles is carried out followed by ligation to a counter oligo carrying a SNP specific address tag plus another universal amplification handle. The ligated products are then amplified using three universal primers: two labeled with fluorescent dyes matching the two allele specific probes and a third complementary to the amplification handle of the address tag probe [117]. The capacity of this assay is currently 3,072 SNP's per reaction and 96 reactions (samples) in parallel generating almost 300,000 genotypes per run.

Another method for multiplex amplification is **trinucleotide threading (TnT)** which shares the end-joining strategy with the MIP's, but by restricting the extension to contain only three out of the four nucleotides, an extra requirement to be fulfilled in order for end-joining to occur is introduced [118]. This method has been applied within SNP-genotyping, targeted

expression profiling and STR genotyping [119–121].

Emulsion PCR

PCR does not provide clonal amplification since its product is a mixture of copies from a variety of starting molecules. Nevertheless, PCR can be used for clonal amplification by compartmentalizing the reaction into microscopic reactors containing only one starting molecule through the creation of a water-oil emulsion [122]. Emulsion PCR (emPCR) is heavily used prior to massive sequencing within many of the leading sequencing platforms today (such as 454, SOLiD and Ion Torrent). Another application of emPCR is amplification of complex libraries, like whole genomes, using universal amplification handles ligated to the template molecules. This can be done using conventional PCR as well but with the risk of introducing bias by size discrimination and product recombination [123].

2.2 DNA sequencing

The process of determining the sequential order of the nucleotides that constitutes a DNA molecule is called sequencing. During recent years, the development of sequencing machines regarding throughput and per-base cost has proceeded beyond Moore's law, a comparison often used for describing something developing at exponential rate. In 1965 Gordon E. Moore, cofunder of the company *Intel*, predicted that the number of transistors fitted on a integrated circuit will double every two year (in the paper from 1965, he actually stated one year but later revised it to two years) [124].

In figure 2.1 the development of sequencing costs at a major genome center since 2001 is shown. Until the end of 2007, the development follows Moore's law but between Oct-07 and Jan-08 this trend was broken. Until then, the sequence data was generated with capillary electrophoresis instruments but when the next generation sequencing instruments entered the market, the development accelerated. The observant reader notices the slightly increase between Jul-12 and Oct-12 and this is probably due to lack of machine performance upgrades and increased costs related to personnel, administration and consumables.

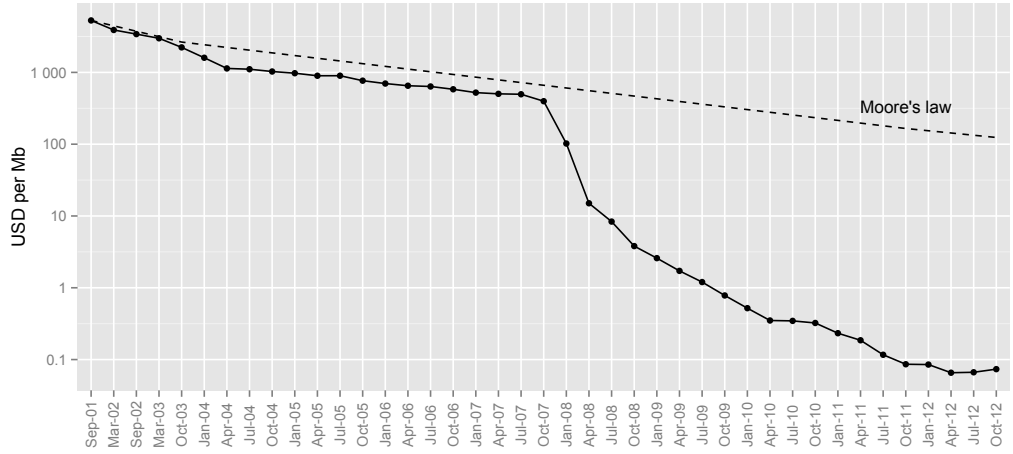

Figure 2.1 The development of sequencing costs in comparison to Moore's law. Data obtained from [125].

Electrophoresis

The history of sequencing starts long before 2001 when the first draft of the human genome was released and even before 1977 when the fist effective sequencing methods and the first genome were published. It actually starts in 1937 when Arne Tiselius, a chemist stationed in Uppsala, used electrical current to separate the proteins within a blood sample in a device he called *electrophoresis apparatus* [126]. In 1952, Markham and Smith used the method to separate digested RNA into distinct bands revealing the length of the molecules [127] and after some improvements by others, such as using a starch or polyacrylamide gel [128, 129], size separation by electrophoresis is one of the most frequent used methods for analyzing nucleic acids.

The principle of nucleic acid electrophoresis is to take advantage of the negatively charged backbone. The sample is loaded in a well on a gel and an electrical current is applied making the negatively charged nucleic acids migrate towards the positively charged terminal. The size of the molecule determines how fast it migrates so that short molecules will travel a longer distance than long molecules within a specific time. Factors to vary for affecting the migration speed of a sample is the density of the gel, where molecules will move slower in a more dense matrix, and the electric field intensity (determined by the applied voltage and the distance between the electrodes, unit: V/m)

where a higher intensity makes the molecules migrate faster. The most common way of visualizing the migrated DNA is by staining the gel using a dye which binds to the DNA molecules and fluoresce when exposed to ultraviolet light. Preferable, the fluoresce is highly intensified when the dye is bound to a DNA molecule compared to unbound dye.

1977 and beyond

Since electrophoresis was the most delicate way of analyzing nucleic acids, a lot of scientists must have deliberated on how it could be used for resolving the mutual order of the four DNA bases that constitutes a DNA polymer. The molecular tools for DNA manipulation available in the mid-70's were basically what we have access to today, such as restriction enzymes, ligases, exonuclease and DNA polymerases (see table 2.1). The key was to generate DNA fragments of various length coupled to the base composition so that they upon size separation could reveal the DNA sequence. Fredrik Sanger

Enzyme	**Action on DNA**	**Year**	**Reference**
Polymerases	polymerization	1956	[130]
Exonucleases	degradation	1960	[131]
Restriction enzymes	cleavage	1965	[132]
Ligases	joining	1967	[133–136]

Table 2.1 The most important components of the molecular toolbox for DNA manipulation and their year of discovery.

presented a method for achieving this already in 1975 denoted "*plus and minus sequencing*". The idea was to generate fragments of all possible lengths by letting a polymerase extend a growing DNA strand from a primer under suboptimal conditions so that extension was very slow and asynchronous. The product was then divided into 2x4 parallel reactions omitting one of the four bases in one set of reactions (minus) and having only one of the four bases present in the other set of reactions (plus). This disrupts the extension at different sites depending on the bases present in the specific reactions and generates molecules of all possible discrete lengths. Size separation by polyacrylamide gel electrophoresis (PAGE) revealed the sequence of the template molecule [137]. The introduction of PAGE was a key for allowing sufficient

resolution for separating longer fragments and achieving longer read lengths (about 50 bases) and Sanger used this method for sequencing the fist genome, which was that of the bacteriophage *phi X174* consisting of 5,375 bases [138].

In 1977, two new methods for DNA sequencing were presented. Gilbert and Maxam solved the puzzle without the use of enzymes by employing chemical treatment for cleavage of the backbone at base specific sites (even thou enzymes may have been used in the labeling process and for generating the template molecules). By dividing a sample of a 5'-radioactively labeled DNA strand into four reactions of different chemical environments restricting cleavage to A, A+G, C and C+T respectively, they generated fragments of various distance between the 5'-probe and the cleavage point. After size separation by gel electrophoresis, the sequence of the molecule could be revealed [139]. This way of sequencing was preferred before Sanger's plus and minus method since Sanger's approach was more tedious and could not handle homopolymers in a satisfactory way [140].

However, these drawbacks were circumvented in Sanger's second method, published in december 1977, in which the chain terminating function was trusted upon dideoxynucleotide triphosphates (ddNTP), nucleotide analogs which are incorporated to the growing DNA chain as normal nucleotides but hinders further extension. By splitting the template sample into four reactions, each of which containing all four native nucleotides and a small amount of one the four ddNTP's, fragments of all possible discrete lengths are generated and the read-out can be done using PAGE [141]. The introduction of ddNTP's as chain terminators allowed even longer reads (about 100-200 bases) and individual read-outs for all residues of homopolymeric regions. This method, known as *Sanger sequencing*, have been widely used and the improvements regarding read length and throughput done in the 80's, such as using fluoresce labeled primers or ddNTP's and automated capillary electrophoresis [140, 142], made sequencing the human genome feasible. Today, Sanger sequencing is the golden standard of sequencing methods and with its high accuracy (>99,9%) it is often used to confirm findings from other more high throughput methods.

Pyrosequencing

During his post-doc period in Cambridge 1986, the swedish chemist Pål Nyrén came up with the idea of sequencing by following the activity of a nucleotide incorporating polymerase in real-time by analyzing its release of pyrophosphate (PPi) [143]. This could be done using a series of enzymatic reactions for converting the released PPi into light and by adding one nucleotide at a time in a cyclic manner while observing whether light was emitted or not, the sequence of the growing DNA chain could be deciphered. It took Pål ten years to develop a working protocol for pyrosequencing, much because of funding issues. Finally, the reaction included four enzymes and steps: i) a **DNA polymerase** incorporates a nucleotide and releases PPi which is used by ii) **ATP sulfurylase** for converting adenosine phosphosulphate (APS) into adenosine triphosphate (ATP) which drives the light emitting oxidation of luciferin into oxyluciferin conducted by iii) **luciferase**, an enzyme isolated from fire flies. Finally, all remaining nucleotides and ATP is degraded by iv) **apyrase** before the next cycle is initiated. The intensity of the emitted light during a cycle is partial proportional to the amount of accessible PPi which makes homopolymeric regions detectable up to a certain limit [144, 145].

The main factors limiting the read length of pyrosequencing are incomplete incorporation of nucleotides to all template molecules which causes some of them to fall behind (minus frame-shift) and incomplete degradation of the remaining nucleotides resulting in some template molecules having too many incorporations in the next cycle and thus gets ahead (plus frame-shift). Foreach cycle, more and more molecules will come out of frame causing higher noise and lower signals. In addition to this, accumulation of the by-products sulphate and oxyluciferin inhibits the activities of ATP sulfurylase and luciferase making the pathway from PPi to light slower with the effect of a longer lasting light signal at lower intensity. The nucleotide degrading enzyme apyrase is inhibited by its product deoxynucleotide monophosphate (dNMP) which also causes the DNA polymerase to work slower and the decreased activity of both these enzymes results in increased frame-shift [146].

Even though pyrosequencing does not perform as well as Sanger sequencing regarding read length and throughput, it is less labour intensive with shorter turnaround time and has found its use for several applications such

as SNP analysis [147], forensic mitochondrial DNA typing [148] and methylation analysis [149].

2.3 Next generation sequencing

As seen in figure 2.1, the sequencing market took a new turn during the second half of the 00's. The keys to this paradigm shift were miniaturization of reaction volumes and parallelization. Before this, sequencing was conducted in microliter reactors, 96 in parallel, but now it became possible to run millions or even billions of parallel reactions in tiny reaction vessels. At present, there are four sequencing platforms on the market: **454**, **SOLiD**, **Illumina** and **Ion Torrent**. Since the release of the ABI 370A DNA sequencer in 1986 [140], the sequencing market was dominated by Applied Biosystems and their capillary sequencing machines and when these massive parallel sequencing technologies were announced, they became denoted *"Next Generation Sequencing"*. Even if these technologies today are the present generation, they have become synonymous to the term "next generation" and some have argued to stop using it. But since everybody in the field know what is means, there is no reason to hinder its use. Figure 2.2 shows the dramatic increase in using next generation sequencing.

A common feature of all these technologies is that they rely on clonal amplification prior to sequencing in order to achieve sufficient strong signals for base calling. To enable this amplification, universal amplification handles need to be incorporated at both ends of the DNA molecules to be sequenced, a process commonly referred to as *library preparation* (see section 2.4). Another feature, common for the three former platforms, is the use of an optical sensor for detecting light generated by various sequencing chemistries.

454

The first instrument to enter the market of massive parallel sequencing platforms was the *Genome Sequencer 20* (GS20) by 454 Life Sciences™ which was released in 2005. By performing pyrosequencing optimized for solid support and picolitre-scale volumes, the instrument was capable of generating 300,000 reads of 100 bases in each run [150]. Since then, the platform has been upgraded several times enhancing both read length, error rates and

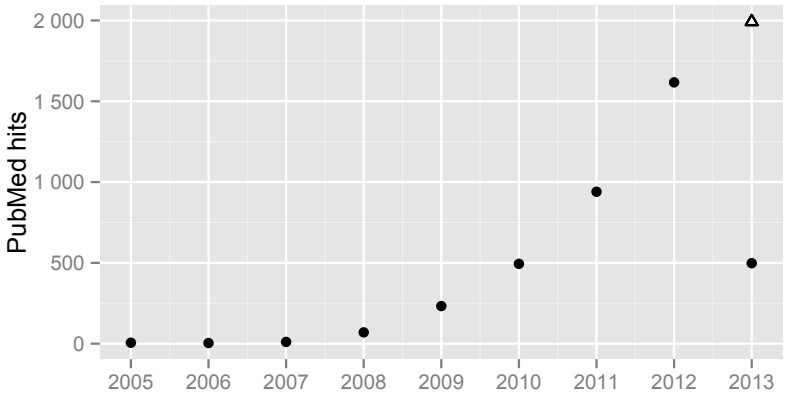

Figure 2.2 Number of hits per year in PubMed from the search string: *"next generation sequencing"*. The triangular point of 2013 is an estimate based on the number of 2013 hits so far.

throughput. The latest version, named *GS FLX Titanium XL+*, is capable of generating one million reads with a length of up to 1,000 bases (mean 700 bases) in a 23 hour run [151]. In 2009, the GS FLX+ system got a smaller sibling when the bench-top instrument *GS Junior System* was released and today it is capable of producing 100,000 reads with an average length of 400 bases in a 10 hour run [152].

As mentioned, 454 uses pyrosequencing chemistry. The genomic DNA to be sequenced is fragmentized through nebulization, end-polished, phosphorylated and adaptors are ligated to the ends. The library is then immobilized onto microscopic magnetic beads for clonal amplification by emPCR. Next, the emulsion is broken and the DNA carrying beads are collected and spread on a *PicoTiterPlate* which consists of millions of picolitre sized wells just big enough to fit one single bead [153]. The bottom of the plate is transparent, allowing a charge-coupled device (CCD) image sensor recording the light emissions.

454 has been able to solve many of the limiting factors concerned with pyrosequencing. The accumulation of enzyme inhibiting byproducts is eliminated since the reaction is performed in a fluidics system which facilitates washing procedures between the cycles. In order to save reagents, the enzymes

luciferase and ATP sulphurylase are immobilized onto even smaller beads, embedding the DNA covered ones so that they can remain in the plate during the washes. The DNA polymerase is not washed away because it is bound to the primed DNA templates. These improvements have contributed to the enhanced read lengths that the 454 is capable of compared to conventional pyrosequecing, which was conducted in standard reaction wells.

The major sequencing errors in the 454 sequencing platform are indels, especially in homopolymeric regions [154]. This is because the light signal produced from the enzymatic pathway, when bases are incorporated, is not perfectly linear to the amount of available PPi and the longer stretch of equal bases, the harder the signal is to be interpreted by the base-calling algorithm. The long reads and relatively short run duration is 454's advantage over the other platforms, especially for assembling low-complex regions when sequencing genomes *de novo* [155] and in metagenomics [156]. However, the read lengths of the competing platforms are increasing and their introduction of table-top sequencers capable of quick runs will make it difficult for 454 to live long and prosper.

Illumina

> *"... But they were all of them deceived for another method was made. In the land of Massachusetts, in the labs of Cambridge Chemistry Department, the scientists Shankar Balasubramanian and David Klenerman developed in secret a master method, to control all others and into this method, they poured their geniality, their illumination, and their will to dominate all other sequencing methods. One method to rule them all."*
>
> *adapted from J.R.R. Tolkien, 1954*

In the summer of 1997, the two colleagues Shankar Balasubramanian and David Klenerman discussed how the fluorescently labeled nucleotides they were working with at the time could be used for massive parallel sequencing of short reads on a clonal array. They worked at Cambridge University and were inspired by the achievements within the filed of DNA research that

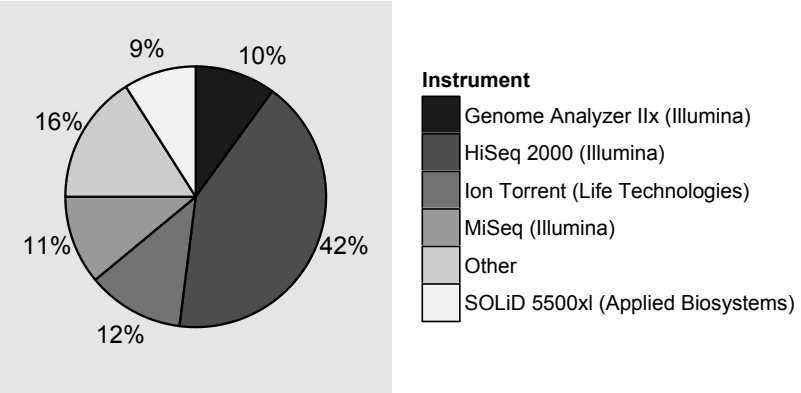

Figure 2.3 The sequencing market shares in October 2012 according to a reader survey done by *In Sequence*. Data obtained from [157].

had been accomplished previously by people like Alexander Todd[1], James Watson[2], Francis Crick[2], and Frederick Sanger[3]. After receiving initial funding, the journey begun: they formed the company **Solexa** in 1998, acquired Manteia's molecular clustering technology in 2004, sequenced the genome of phiX-174 (same bacteriophage as Sanger in 1977) in 2005, launched their first instrument in 2006, was acquired by Illumina in early 2007 and sequenced the human genome in 2008 [157, 158].

Today, Illumina is by far the most widespread sequencing platform and according to a reader survey by *In Sequence* performed in october 2012, they hold 66% of the sequencing market (see figure 2.3) [159]. The method, commonly referred to as *sequencing using reversible terminators*, uses fluorescently labeled nucleotides which are blocked for further extension and hence, only allows for one nucleotide being incorporated per cycle. But prior to

[1]Worked with synthesis and structure of nucleotides, awarded with the Nobel Prize in Chemistry 1957.

[2]Co-discoverer of the structure of DNA, awarded with the Nobel Prize in Physiology or Medicine 1962.

[3]Inventor of the Sanger sequencing method and determined the structure of various proteins, insulin among others. Was awarded the Nobel Prize in Chemistry 1958 for the latter and the Nobel Prize in Chemistry 1980 for the former.

the actual sequencing reaction, the DNA sample must be processed into a sequencing library and clonally amplified. The library preparation includes random fragmentation, end-polishing, phosphorylation and adenylation followed by adaptor ligation and enrichment of ligated fragments by PCR (see chapter 2.4 for details). In this process, sample specific identification tags are incorporated so that several samples can be sequenced together but still have separate data files.

After quantification, the sequencing library is diluted to an appropriate concentration, made single stranded by adding NaOH and spread on the flow-cell surface for formation of a clonal single molecule array. The flow-cell is covered with surface bound oligonucleotides complementary to the ligated adapters which allows the single stranded DNA to anneal. The surface bound oligonucleotide primes an extension reaction forming a copy of the template molecule covalently attached to the flow-cell surface and the original strand is then denatured and washed away. The adapter sequence at the free end of the copied strand is then annealed to another complementary surface bound oligonucleotide which primes another extension reaction forming a bridge. When the product is denatured, both ssDNA molecules remains on the surface and can form bridges for another cycle of amplification. Thermal cyclic reaction conditions allows for several rounds of denaturing, annealing and extension which generates clusters of molecular copies [158, 160]. In the Illmunina sequencing platform, the bridge amplification is managed by a so called *cBot* which is an automated system for cluster generation.

The flow-cell is then transferred to the HiSeq instrument for sequencing by synthesis. In contrast to pyrosequencing, where only one nucleotide is present in each cycle, all four nucleotides are subjected to incorporation in sequencing using reversible terminators. This is made possible thanks to the base specific fluorescent label which facilitates identification of the incorporated nucleotide. Since each nucleotide carries a blocking group bound to the 3'-carbon of the sugar unit, only one nucleotide is incorporated in each cycle and the incorporation reaction can be driven to completion without risk of over-incorporation [158]. Once the incorporation is complete, the base is determined by fluorophore excitation by laser and CCD camera imaging. Prior to the next cycle, the blocking group and the fluorophore are chemically cleaved off from the nucleotide leaving a 3'-hydroxyl group suitable for

another incorporation. Once the read is completed, usually after 101 bases, the product is denatured and removed and the remaining strand is used for generation of another bridge and the template for a second read is synthesized.

The main reason for low quality reads is over-clustering which means that too many ssDNA molecules were spread on the flow-cell surface generating mixed and overlapping clusters that are hard to interpret for the base-calling algorithm. The most abundant sequencing errors are substitutions and specific sequences, such as inverted repeats and GGC regions, have been shown to be more prone to errors [161]. The suggested mechanism behind the sequence specific errors is secondary structure formation from base pairing within the ssDNA molecule resulting in a hair-pin structure that blocks the DNA polymerase and hinders complete nucleotide incorporation.

Currently, Illumina has five sequencers in their program (see table 2.2) and provides kits for several kinds of applications like whole-genome sequencing, targeted sequencing, RNA-sequencing, methylation analysis by bisulfite sequencing and protein-DNA interaction analysis by ChIP-seq. The combination of medium read lengths, high quality paired-end reads and a massive throughput has made Illumina the system of choice for a vast majority of the sequencing project conducted around the world.

	HiSeq 2500/1500	**HiSeq 2000/1000**	**MiSeq**	**GAIIx**	**HiScanSQ**
Output (Gb)	600/300	600/300	8.5	95	150
Run time	27h-11 days	8.5-11 days	4-39h	14 days	8.5 days
Single reads (Billion)	0.3-3	3-6	0.017	0.32	0.75
Read length (bp)	100-150	100	250	150	100

Table 2.2 Illumina's current line-up of sequencers as of March 2013.

SOLiD

The 3rd platform to hit the next generation sequencing market was released in 2007 and is based on sequencing by oligonucleotide ligation and detection (SOLiD). This is the only commercially available sequencing platform today that does not include a DNA polymerase in its sequencing reaction

but instead relies on a DNA ligase [162]. However, the introduction of universal adaptors through library preparation and a clonal amplification by emPCR prior to the sequencing is shared with the others. The sequencing method, originally denoted *polony sequencing*, was developed at Harvard Medical School in Boston by Greg Porreca and Jay Shendure under supervision of George Church during the mid-00's and was later adapted to the SOLiD platform by the company Applied Biosystems.

The sequencing reaction is conducted on the microscopic magnetic beads used in the emPCR that are covalently bound to the surface of a glass slide through 3'-modifications of the emPCR products. The sequencing reaction starts by hybridizing a randomized DNA probe next to a sequencing primer. If the two bases closest to the primer match the template DNA strand, the probe can be ligated to the primer and remaining non-matching probes are washed away. The probe has a fluorescent label specific to the two discriminating bases that is recorded before half the probe, including the fluorophore, is cleaved off for another round of sequencing. Since there are two discriminating bases, SOLiD uses a two-base encoding system consisting of 16 different base combinations labeled with four different colors and in order to decipher a base, the combined color information of two probes is used. This basically means that each base needs to be sequenced twice in order to be identified but on the other hand, this system gives a very low error rate. Since a ligation reaction is used, sequencing can be done in both directions of the primer which is not possible when using a DNA polymerase [163].

In the newest instrument, the *5500 W Series Genetic Analysis System*, another approach for clonal amplification prior to sequencing is offered. Through a isothermal process, the DNA library is amplified directly on the FlowChip. Like in bridge amplification, oligos that are attached to the surface functions as primers for annealed template molecules. After extension, instead of annealing the 3'-end of the newly synthesized strand to a new surface bound primer, the template strand switches primer from the extended to a non-extended. At the same time, a free primer is added to the 3'-end of the newly synthesized strand so that both strands of the first product is copied in the second cycle. This method, called *Wildfire template walking*, produces colonies suitable for sequencing and allows for a more efficient sequencing reaction that consumes less reagents and a higher density on the sequencing

chip resulting in more data at lower cost. This system is capable of producing 320 Gb of data in a single run from 2x50 bp reads [164].

Ion Torrent

The latest addition to the sequencer market was the Ion Torrent platform, released in 2011. This is the only commercially available sequencing technique today that uses neither light nor optics in its base detection system. Instead, a semi-conductor chip is used to detect the alteration in pH following a proton (H^+) release when a base is incorporated [165]. Like in pyrosequencing, the method relies on subjecting natural nucleotides to incorporation, one base at a time in a cyclic manner, and then detect if incorporation took place or not. This means that several nucleotides can be incorporated within the same cycle in homopolymeric regions and because of this, Ion Torrent has inherited its predisposition of indel errors from 454 with an accuracy of 96.5% calling a 5-mer. However, the substitution error rate is less than 0.1% and the mean accuracy of a 250-bp read is 99.7% [166]. The fact that the sequencing chemistry entirely consists of natural reagents, enzymes as well as nucleotides, the enzyme activity is kept at its best and the reagent cost is low. Also the detection chip is cheap since it is based on CMOS processes from standard electrical component production plants.

There are three different chips available for the Ion PGM (Personal Genome Machine) sequencer named 314, 316 and 318. The different chips have different throughput and there are also several read-length options which gives a multitude of choices when designing a sequencing experiment. The scalability ranges from generating 3 Mb of data in 30 minutes using the 314 chip and 35 bp reads to 2 Gb of data using the 318 chip and 400 bp reads in 7 hours and 21 minutes. This level of throughput makes the machine ideal for applications like amplicon sequencing, SNP confirmation and small genome sequencing. Another machine called the *Ion Proton System* was recently made available for featured costumers and is currently capable of generating 10 Gb of data in a 2-4 hour run. Nevertheless, an upcoming chip is promised to offer whole human genome sequencing at 20× coverage, from sample to called variants, in less than a single day.

In conformity with other platforms, a library must be prepared and subjected

to clonal amplification using emPCR prior to sequencing. Multiple copies of the template in each reactor is needed so that the amount of released H^+ can reach detectable levels. However, emPCR is template consuming, expensive and technically challenging so Ion Torrent have announced that they are developing a non-emulsion PCR protocol for producing templated beads called *Avalanche*. This is based on local amplification on beads but without the emulsion which is achieved by having one primer bound to beads and the other in solution [166].

2.4 Preparing DNA for massive sequencing

The science-fiction movie *"GATACCA"* from 1997 sets in a not too distant future where all humans are categorized by their genetic make-up. In one scene, the female main character (Uma Thurman) visits a booth on the street, where whole-genome sequencing and genetic make-up scoring is conducted, for investigating a hair she just collected at the male main character's (Ethan Hawke) working desk. The dialogue between her and the man behind the desk goes like:

> - *You want the full sequence?*
> - *Yes.*
> 20 seconds later he gives her a thick paper roll and reads from the first line:
> - *9.3. Quite a catch.*
> - *Yes, quite a catch.*

So, he just performed whole-genome sequencing, data analysis and risk assessment in 20 seconds (not mentioning printing all 3 billion characters on a piece of paper). Currently, we are not there (yet).

Today, this would take weeks and require a lot more DNA than what is in a hair. The process from biological sample to interpreted sequence data can be subdivided into five steps: i) DNA extraction, ii) library preparation, iii) clonal amplification, iv) sequencing and v) data analysis. Library preparation is the process where the DNA molecules are adapted for the sequencing reaction by adjusting their length and adding synthetic sequences, commonly denoted adapters, to their ends. In all next generation sequencing platforms available today, library preparation starts with **fragmentation** of

the DNA sample in order to generate short randomly sheared molecules and this is followed by **modification** of the DNA fragments adapting them for the sequencing reaction and finally **quantification** of the produced library. As the amount of data generated by the sequencing machines have increased and the sequencing costs have fallen, the library preparation step has become a bottleneck for the sample throughput and has a relatively high impact on the budget of a sequencing project.

Fragmentation

The concept of sequencing randomly sheared fragments is commonly denoted *shotgun sequencing* and the term was used as early as 1979 when a paper described the adaptation of computers for assembling shotgun reads into long continuous sequences [167]. There are three ways of shearing DNA: by physical stress, enzymatic cleavage or chemical treatment. The most common method in library preparation procedures is physical shearing by sonication. Here, DNA is subjected to ultra-sonic sound waves in a sonicator and the company *Covaris* provides an instrument where this can be done in small vials and in downstream processing friendly solutions with relatively small material loss. Other methods of physical shearing is nebulization that uses gas pressure (used by 454) and hydroshearing that uses change in fluidic pressure.

Enzymatic approaches relies on several enzymes which together cleaves the DNA at random positions. *New England Biolabs* (NEB) sells a product called *dsDNA Fragmentase* which is a mixture of two enzymes, one that randomly nicks the dsDNA and one that recognizes the nicks and cleaves the opposite strand. The sizes of the produced fragments are dependent of the reaction duration. One advantage of using enzymatic shearing is that is is easy to scale and automatize but enzymes requires the right buffer to work optimally and this is not always compatible with downstream reactions and a purification step might be needed in which some of the material is lost.

Modification

The purpose of the modification step is to incorporate the sequencing adapters to the DNA molecules and to enrich for ligated fragments. But before liga-

tion of sequencing adaptors can be made, the ends of the fragmentized DNA molecules have to be repaired and commonly, also adenylated and phosphorylated. All these end-modifications are performed by enzymes that have their own optimal reaction environment and temperature. But since all purification steps involves loss of material, it is preferable to have as many enzymes as possible within each reaction but without decreasing their reactivity by deviating from optimal reaction conditions. This is further investigated in **paper III** and is also discussed in chapter 3.3.

As mentioned, the ends of the fragmented DNA needs to be end-repaired since they might have large 3'- or 5'-overhangs unsuitable for adaptor ligation. This is commonly done using the enzyme *T4 DNA polymerase* which fills the 5'-overhangs by extending from the innermost 3'-end and it also exhibits 3'-exonuclease activity and degrades 3'-overhangs. This results in blunt-ended DNA fragments which can be directly subjected to blunt-end ligation. In order to ligate the 3'-end of one DNA strand to the 5'-end of another strand, the 5'-end needs to be phosphorylated. For adaptor ligation, one can have pre-phosphorylated adapters or phosphorylate the end-repaired DNA, commonly using the enzyme *T4 polynucleotide kinase* (PNK) or both. It is desirable to minimize adaptor-adaptor ligations and also ligation of two end-repaired fragments. The former occurs in greater extent than the latter since the adapters are short and often present in ten-fold molar excess. There are several ways of avoiding this of which one is to use non-phosphorylated adapters which cannot be ligated to each other. Another way is to use so called *AT-cloning* where the end-polished DNA is adenylated which means that one A-base i added to the 3'-end. This can be done using a DNA polymerase lacking 3'->5' exonuclease activity such as *Klenow fragment*. Only adapters having one complementary T-base at their 3'-end can then be ligated to the A-tailed fragments but neither adapters nor A-tailed fragments can be ligated to each other.

Once the fragmented DNA is prepared for ligation the adapters can be added. When two kinds of adapters are used, here denoted A and B, in order to incorporate different synthetic DNA sequences at the different ends of the DNA fragments, there are three possible outcomes of adapter combinations: fragments carrying A and B, which is the desired, but also A-A and B-B carrying fragments will be formed. This means that two thirds of the formed fragments

will be useless and the yield from the ligation reaction will be low. However, this can be circumvented, while still having different sequences added to the ends, by ligating a Y-shaped adapter which is a adapter construct that is complementary in one end (the base of the Y) but have different sequences in the other end (the arms of the Y) [168].

After the ligation step, the DNA is basically ready for sequencing but since the different enzymatic steps do not have 100% efficiency, only a minor fraction of all fragmented DNA that was put in the modification process have ended as complete library molecules. It is therefore desirable that the ligated fragments are enriched and amplified by running PCR using primers complementary to the ligated adapters. It is also common that barcodes, enabling multiplex sequencing of several samples within the same lane, are incorporated in this step through one of the PCR-primers. However, GC-rich DNA-fragments are discriminated in the PCR-reaction since it is harder for the DNA polymerase to extend through these regions and hence, bias is introduced. To overcome this, methods for amplification-free library preparation have been developed where the discrimination between ligated and non-ligated fragments is done on the flow-cell surface during cluster generation instead, but more input-DNA is required in order to generate sufficient ligated molecules [169, 170]. Another problem introduced in the PCR-step is the formation of PCR duplicates. If the PCR is allowed to run for too many cycles, these will dominate the library and the sequencing data will be mostly redundant.

Quantification

In order to load an appropriate amount of molecules to the clonal amplification procedure (cluster generation or emPCR), it is important to quantify the generated library correctly. In the Illumina instruments, the data quality is highly dependent on an optimal cluster density since too dense clusters will grow into each other, generating mixed signals, and too sparse clusters will generate less data. DNA amounts can be measured by binding fluorescent dyes to the DNA molecules and measure the fluorescence or by assessing the absorbance using a spectrophotometer. The former is more accurate than the latter since absorbance based approaches are more sensitive to contaminants. Both methods suffer from measuring the total amount of DNA in the

sample with no possibility of discriminating between ligated and non-ligated fragments nor primer-dimers and other short fragments [171]. Automated capillary electrophoresis, like the *Bioanalyzer* by Applied Biosystems, offers the possibility of analyzing both the concentration and size distribution of the library. But in order to measure the exact amount of ligated fragments one has to rely on a PCR-based approach such as quantitative PCR (qPCR) in which only amplifiable molecules will generate a signal.

Library complexity

The effectiveness of the library preparation process and the amount of starting material determines the complexity of the produced sequencing library. Library complexity is defined as *"the expected number of distinct molecules that can be observed in a given set of sequenced reads"* [172]. Figure 2.4 illustrates a high complex library containing many kinds of point characters and a low complex library containing an equal amount of points but only two kinds of characters. In analog, a highly complex sequencing library contains many unique kinds of randomly sheared, modified and amplified molecules whilst a low complex library contains few unique molecules that have been copied through PCR so that the mass of the libraries are equal. If the modification step of the library preparation process is inefficient, a relatively small amount of the starting material will be available for amplification in the PCR-step and to have sufficient molecules for the sequencing reaction, additional PCR-cycles have to be run and the complexity of the library would dwindle.

Library complexity is hard to assess prior to sequencing since the quality controls available are mainly quantitative but shallow sequencing prior to a large run can give some indication [172]. When interrogating sequence data, low-complex samples are shown as having a high proportion of PCR duplicates. Aligned reads sharing genomic start and stop coordinates have likely originated from the same template molecule and thanks to this, they can be computationally identified and are often filtered away to avoid redundant data. So, the proportion of PCR duplicates in a sequence data set depends on the complexity of the sequenced library but it also determined by the sequencing depth.

Consider the two libraries illustrated in figure 2.4. In a sequencing experi-

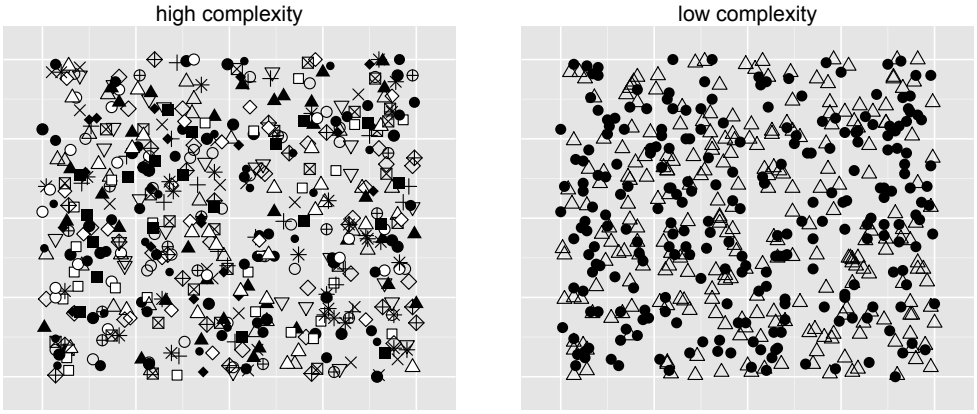

Figure 2.4 Illustration of high and low complexity sequencing libraries.

ment, reads are drawn randomly from a library of molecules so imagine that the points that illustrates molecules in the figure are randomly sampled. How many points will have to be drawn before two equal characters are likely to be observed? In the low-complex library there are only two different characters so at the second attempt there is a 50% chance that the new character is of the same kind as the first (assuming they are present in equal quantitates) and when three points are drawn there is certain that at least two of these are of the same character. This means that at a shallow depth of one and two drawn points, the library can still have no duplicates but after a few extra points, both kinds of characters present in the library are likely to have been observed and all extra points are abundant. The high-complex library consists of 20 different point characters so not until the 21st attempt there is certain that two equal characters have been drawn.

In figure 2.5, two real libraries, one of high complexity and one of low complexity, have been sequenced at a depth of more than 80 million reads. To illustrate the sequence depth dependence on the proportion PCR duplicates, these data sets have been subsampled at various depth and the proportion duplicates have been calculated and plotted. Since the probability of observing new unique molecules decreases when more and more molecules are observed, the gain of obtaining more reads becomes less when sequence depth increases. As shown in figure 2.5, the number of unique molecules observed

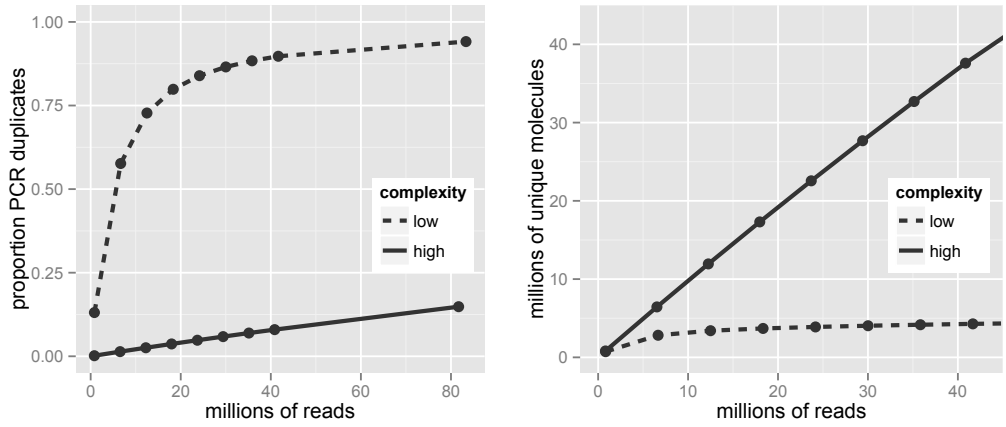

Figure 2.5 *LEFT*: the sequence depth dependence of observed proportion PCR duplicates in high and low complexity sequencing libraries. *RIGHT*: the increase of observed unique molecules at higher depth is almost linear in the complex library but is early saturated in the low-complex library.

becomes saturated already at a little more than ten million reads in the low-complex library whiles the complex library almost show a linear relation between sequence depth and observed unique molecules.

Complex sequencing libraries are highly preferable and are achieved by using large quantitates (micrograms) of starting material when preparing the library. However, this is not always feasible, especially when working with biological material coming from limited sources such as blood samples, tumor tissue or biopsies. Under these circumstances, it is most desirable to get as high complexity as possible. One way is to apply whole genome amplification before library preparation but this technique comes with the risk of introducing bias. Another way is by having a high yield library preparation procedure. One of the most effective methods that are available today uses an *in vitro* transposition procedure instead of the traditional fragmentation, modification and ligation approach. In this method, a transposase enzyme carries out both fragmentation and adapter ligation in one single step which reduces the amount of input DNA needed down to as low as 100 pg [173]. This technology is now sold by Illumina under the name *Nextera* and according to its standard protocol, 50 ng DNA is sufficient for producing a highly com-

plex library in 1.5 hours with only 15 min hands-on time. Another method that uses the more traditionally enzymatic steps is the *ThruPLEX-kit* sold by Rubicon Genomics. Here, a stem-loop adapter construct is blunt-end ligated to the fragmented and end-repaired DNA molecules with high efficiency and since its 5'-ends are blocked, the formation of adaptor-adaptor ligated molecules is reduced. After ligation, the left-over adaptors are degraded and amplification can be performed without any purification step.

Barcoding

Since the great capacity of todays sequencing instruments often exceeds whats needed to answer the biological question behind a sequencing experiment, the ability of analyzing several samples in the same run, but still be able of analyzing the data as discrete samples, has proven to be of importance. This can be achieved by introducing a molecular sample specific barcode, which consists of a short sequence stretch situated next to the sequencing primer region. The introduction can be done either through PCR [174, 175] or ligation [176]. A method for using a combination of two tags when sequencing large sample sizes has been developed in **paper I** and is further discussed in chapter 3.1.

2.5 Targeted sequencing

Experiments where only selected parts of genomes are analyzed are called targeted sequencing. This is preferable when the biological question behind the experiment does not require investigation of the whole genome but only a pre-defined part. This sub-part can consist of for example a set of genes involved in a biological pathway, all coding regions (whole exome), genes previously associated with a trait in a GWAS or protein binding regions. There are several approaches for enriching the desired regions but the most used commercial methods are based on either PCR or hybridization.

At first, hybridization selection was performed on a chip surface [177, 178] with relatively low efficiency and complicated laboratory procedures but when in-solution hybridization was presented [179], it became both better and more accessible. In solution hybridization based approaches, sheared and adapter ligated DNA fragments are allowed to hybridize to biotinylated

probes of synthetic DNA or RNA, complementary to the target regions, often denoted *"baits"*. The hybridization reaction is often allowed to run for several days so that the molecules of the highly complex sample should have time to find and hybridize to their complementary baits among all other bait molecules in the highly complex bait library. Once the hybridization reaction is completed, the target carrying bait molecules are bound to para-magnetic beads through a biotin-streptavidin coupling and non-target molecules can be washed away. Target molecules are then eluted and amplified. If the bait molecules consisted of RNA, like in Agilent's *SureSelect* technology, they can be easily degraded by adding RNAse. The success of the enrichment process can be assessed by runnig qPCR over selected control regions. A commonly used measure of non-uniformity in coverage across the target regions is *fold 80 base penalty*. This estimates the fold extra sequencing needed to raise 80% of the targets above the average coverage. A low value means a uniform sequence coverage and in general, less than 4 is considered good and exomes often ends up around 3.

The three most established technologies for hybrid capture today are *SeqCap* by NimbleGen, *SureSelect* by Agilent and *TruSeq Enrichment* by Illumina. The largest differences between these technologies are the quantity of baits and the density of the bait design. NimbleGen uses a highly dense and overlapping bait design covering the target regions several times, Agilent designs the baits to reside immediately next to each others with 1x target coverage and Illumina uses the most sparse design and relies on paired-end reads to extend the coverage outside the bait regions instead. In a direct comparison of these three technologies's performances in an exome sequencing experiment done in 2011, NimbleGen's dense design (>2.1 million baits) covered least genomic regions (44 Mb) and required the least amount of sequencing to be able to detect SNV's and indels [180]. With additional sequencing however, Agilent and Illumina could detect a greater total of variants (0.66 and 0.34 million baits covering 52 and 62 Mb respectively).

Since NimbleGen performed best regarding both target enrichment efficiency and off-target enrichment (see table 2.3), this is the most cost-efficient technology. Today however, it is the sample preparation including sequence capture that is the most abundant cost and not the actual sequencing (unless the target region is really large) so the best value for money is more dependent

Platform	Target size	Baits	>10x target cov	Off-target
NimbleGen	44 Mb	>2.1 million	96.8%	9.1%
Agilent	52 Mb	0.66 million	89.6%	12.8%
Illumina	62 Mb	0.34 million	90.0%	35.6%

Table 2.3 Comparison of exome sequencing performance using a normalized depth of 80M mapped reads. The comparison was made in 2011, data obtained from [180].

of which per-reaction price that can be negotiated from the respective company. Also, the balance between coverage and efficiency might be considered regarding the demands from the biological question. For example, Illumina's exome design also includes the untranslated regions (UTRs) which might be of interest.

Target enrichment can also be achieved through PCR, designed for both single, hundreds and even thousands of amplicons. In contrast to the hybridization based approaches, PCR performs the discrimination between target and non-target regions directly on genomic DNA and hence, requires less input material. As mentioned in chapter 2.1, it is troublesome to perform multiplex PCR when the number of amplicons increases. In order to amplify a multitude of targets, an approach to prevent primer-dimers from dominating the end-product must be applied. In chapter 2.1, methods for enriching and amplifying SNP's were discussed but there are also several methods for targeting larger regions suitable for sequencing, some of which are kept secret by the manufacturer of business reasons.

Ion Torrent provides a technology called *AmpliSeq* in which a multiplex PCR is performed followed by a primer removal step using a secret reagent that degrades the primers specifically. Since the duration of a run on the Ion PGM is short (4 hours for 400 base reads) and the duration of a hybridization capture is long (1-4 days), the bottleneck of sequence capture using hybridization is inconveniently narrow. Motivated by this, Ion Torrent developed their own capture technology capable of processing a DNA sample into a sequence ready capture library in less than 3 hours. This method allows for amplification of up to 4,000 targets from as little as 10 ng DNA followed by integrated library preparation and sequencing [166]. Ion Torrent

offers both pre-designed panels like the Cancer HotSpot panel, the Inherited Disease Panel and the Comprehensive Cancer Panel but it is also possible to make own custom panels using their on-line design tool.

There are also several methods based on circularization. Agilent offers a technology called *HaloPlex* in which the DNA sample is digested using restriction enzymes and a probe library hybridizes to the specific ends of the target fragments and guides them to form a circle. The probes are biotinylated and allows for target retrieval using streptavidin coated para-magnetic beads. After closing the target circles by ligation, the library is amplified by PCR in a way so that only closed circles becomes amplified [181]. As mentioned in chapter 2.1, padlock probes can also be used for exome capture [115, 116].

2.6 $(n+1)$th generation sequencing

Since the term *next generation* is taken by the present generation, the term $(n+1)$*th generation* may be used instead to address the actual coming generation. $n+1$ is derived from a mathematical method used for proving that a given statement is true (or false) for all natural numbers. It is called *mathematical induction* and it says that if a statement is true for any given number n, it must also be true if n is replaced by $n+1$. In analog with the naming of sequencing generations: if the present generation is generation n, the next one will be generation $(n+1)$.

Today, it is hard to categorize sequencing methods into distinct generations. Some might claim that Ion Torrent belong to another generation than 454, SOLiD and Illumina since there is no light involved in the detection system (the post-light generation) whiles some might claim that Pacific Biosciences (described below) should belong to this generation when regarding its throughput capacity and light based detection. But everyone is right and nobody is wrong and it might be so that the over-nite sensation seen during the late 00s that has led to a paradigm shift in the field of sequencing will not happen again. It is likely that emerging technologies are released when they are still at an early state and then gradually developed over several years before their full potential is reached, even though the emerging nanopore technologies (described below) seem promising from the start.

So, what is to expect from the emerging technologies? While the capacity continues to increase, some new features will also be seen. Among the ones that have already begun to appear are extremely long read lengths and library-free sequencing. In 2009, the Californian company *Pacific Biosciences* (PacBio) showed in a Science paper that they could sequence a DNA strand by fixating a DNA polymerase at the bottom of a zeptoliter (10^{-21} liter) reactor and detect fluorescently labelled nucleotides in real-time as they were being incorporated to the growing strand by the polymerase [182]. The technology is called single-molecule real-time (SMRT) sequencing and it uses an engineered phi29 DNA polymerase immobilized in a zero-mode waveguide (ZMW) which is a nano structure used for reaction confinement and enables detection of nucleotides while they resides at the active site of the polymerase, without the interference from surrounding nucleotides despite their high concentration. The fluorescent dye is attached to the terminal phosphate group of the nucleotide and hence, is naturally cleaved off when the nucleotide is incorporated. In 2011, the company introduced their first instrument, the *PacBio RS*, to the sequencing market and it is today capable of generating 22,000 reads with an average length of 4,600 bases (up to 20,000) in 2 hours. However, the base quality is relatively low with an accuracy of $<$90% at 1x coverage [183]. However, by utilizing the exceptionally long read length, the base quality can be improved by circulating the template to be sequenced and allow the heavily strand displacing polymerase to loop through it several times.

Oxford Nanopore Technologies is a company that develops a sequencing technology based on flowing a single DNA strand through a nanopore and record the shift in current that occur when different bases passes through. The company was formed in 2005 and when they revealed their system at the conference AGBT (Advances in Genome Biology and Technology) in february 2012, the first instrument release was scheduled to late 2012 but at the writing moment (late Mars 2013), no system has yet been made available. There were two instruments announced: the *GridION* and *MinION* systems. The first is a electronics-based platform consisting of several GridION nodes and can be scaled up like computer clusters but one single node can also be used alone like a bench-top sequencer. The MinION applies the same sequencing technique but in a miniaturized disposable device connected di-

rectly to a laptop through a USB-port. According to the announcement, the technology will be able to produce reads of 100 kb at launch and one GridION node will consist of 2,000 nanopores and later extended to 8,000 pores of which each is capable of sequencing hundreds of kb per second. Theoretically, this would allow for sequencing the human genome in 15 minutes using 20 nodes. The technology was demonstrated by sequencing the entire 5.4 kb genome of the famous phage phi X in one single read stretch with an error rate of 4%. However, the company claims the the error rate will be improved until release to around 0.1-1% [184].

Chapter 3

Present Investigations

The papers that constitutes this thesis all aim to overcome technical hurdles arisen from biological questions. The common theme is the aim of making investigation of DNA using sequencing more effective. There are four papers of which two (paper I & II) were conducted at the Royal Institute of Technology (KTH) and two (paper III & IV) at Karolinska Institutet (KI). In this chapter, the projects behind the papers are described and discussed while trying not repeating what is written in the papers.

3.1 Paper I - dog tags

This work was initiated in 2008 when the division of gene technology at KTH recently had gotten their first next generation sequencing instrument in their hands, namely a 454 genome analyzer. This had off course generated a lot of enthusiasm in the group and the ideas of technical methods taking advantage of its high throughput were flowing.

Background

The read length of the 454 instrument makes it suitable for amplicon sequencing since relatively long amplicons can be read through in a single read and its throughput opens up for investigating a large amount of samples at once in the same run. Introducing sample specific molecular barcodes at the ends of the molecules to be sequenced makes it possible to determine the origin of a read and thereby demultiplexing the data by computational methods [174–176]. What is limiting the amount of samples that can be

investigated together is the availability of unique tags. To be able to separate tags containing sequencing errors, it is preferable to design them so that they differ at more than one position and because of 454's predisposition to indel errors in homopolymeric regions, these are also best avoided. It is also preferable to keep the tags short so that they not occupy too much of the reads. When taking these considerations into account, the number of unique 7 bp tags that can be obtained with a minimum of 2 substitutions is 173 and with a minimum of 3 substitutions it is only 52 [176].

Today, Roche provides up to 132 multiplex identifiers (MID's) and if larger sample sets needs to be investigated on the 454, physical separation of samples can be achieved by applying an up to 16 region gasket. This facilitates 2,112 samples to be investigated within the same run but the gasket occupies a significant amount of the sequencing space and preparing 16*132 sequencing libraries using the standard kits is both expensive and laborious.

Several methods circumventing these limitations have been proposed and the overall achievement of these have been to sequence more samples than unique tags used. In other words: addressing the linear increase of needed tags when sample size grows. However, this has been accomplished by the cost of increased complexity of sample handling with higher risk of human errors which may result in erroneous sample tagging and biased data. One such method is denoted *DNA sudoku* and relies on tagging groups of samples rather than single samples and having each specimen present in two unique groups. The groups are then pooled both row-vise and column-vise creating a sample matrix where the total number of pools to be sequenced equals to the number of rows times the number of columns. After sequencing, the matrix can be solved using a computational method similar to the strategy used when solving a sudoku and this makes it possible to associate a genetic diversity to a single sample [185]. This approach greatly decreases the number of tags needed for a certain level of multiplexity but the experimental complexity is large and the method is best suited for finding rare variants within large sample sets.

A smart way of reducing the number of unique tags desired for a certain experiment is by letting the combination of two tags be the key for associating reads to specimen. This was first done in 2010 by a french group that used

tagged PCR-primers to introduce different tags at the two ends of the PCR products so that every sample was given a unique tag combination [186]. Like DNA sudoku, this method was also able to reduce the amount of required unique tags but since all tag-combinations are introduced at the same step, the sample handling and laboratory procedures are still relatively complex. In addition, chimeric molecules may be formed in the library preparation and since there should be different tags at the different ends of the reads, these cannot not be detected and filtered away in the data analysis which may result in reads originating from two different samples being erroneous associated with a third sample.

Aim

If thousands of samples are to be tagged and prepared for sequencing it is desirable that the procedure is as simple as possible so that human errors can be avoided already at the design phase. The aim of this project was to develop an easy scalable dual-tagging method for amplicon sequencing of thousands of samples at a low cost.

Samples

To prove the concept of this method, the 270 bp long 2^{nd} exon of the hyper variable gene DLA-DRB1 was amplified and sequenced for 4,700 samples derived from dogs and wolves originating from all over the world. By comparing the allelic diversity of different geographical regions, the data can then be used for confirming the geographical location of where the domestication of the dog took place described previously [187, 188]. Of the 4,700 samples, 2,059 were collected by buccal swabbing and FTA-cards, 1,830 were blood samples and 819 were hair samples.

Methods

The idea behind the method developed within this project is to use a two-tagging strategy and a smart pooling procedure for reducing both number of unique tags needed and sample handling complexity. This is achieved by introducing the first barcode in the PCR step using 96 tagged PCR primer-pairs, each of which is assigned to a specific well in the 8 x 12 grid that constitutes a traditional 96-well PCR plate. Hence, they are named A1,

A2, ..., H12 and denoted *position tags*. After amplification and tagging of 96 samples, the PCR products are pooled. In order to avoid the time-consuming step of determine the concentrations of all samples, the PCR is cycled relatively many times to ensure that all samples reaches the amplification plateau and hence a somehow equal amount of end-product. This should make the method robust enough for allowing pooling of equal volumes using a multi-pipette rather than setting 96 different volumes per plate for equimolar pooling using a single channel pipette.

All 96 samples from the plate have now become one single sample which is purified from remaining primers and PCR reagents before the second tag, specific for this plate and denoted *plate tag*, is introduced by ligation together with the sequencing adapters. The whole procedure can then be repeated until the desired amount of samples have been processed. In our case, 52 plates were made which gave a total of 4,992 uniquely tagged samples consuming only $96 + 52 = 148$ unique tags. The number of desired tags t for a certain sample size N can be calculated as

$$t = \frac{N}{96} + 96$$

This procedure of tagging and pooling ensures a reliable and easy scalable system since the most complex step, the position-tag PCR, is having the same primer set-up for each plate and hence, a fool-proof procedure of transferring primer-pairs between a primer stock plate and the sample plate using a multi pipette can be set-up. Once the PCR-step is carried out, the number of samples to be handled is almost 100-fold decreased making the remaining steps easy to handle.

In order to further decrease costs and hands-on time in this project, the plate-tag and sequencing adaptor ligation was carried out using an automated protocol in a liquid handling robot [189]. To ensure an even read distribution between the plates, the ligation products are quantified before pooling them equimolar prior to emPCR and sequencing. The demand for unique plate-tags in this particular experiment was further reduced by using a three-region gasket in the sequencing run and hence, the plates were pooled into three sequencing libraries.

Results and discussion

In order to estimate the success rate of the position-tag PCR's, about half of the samples were randomly picked and analyzed using agarose gel electrophoresis or automated capillary gel electrophoresis. This resulted in a success rate estimate of 79% and since samples were added to 4,708 wells (248 wells were negative PCR controlls), the total number of PCR products was estimated to 3,700. After sequencing on a 454 GS FLX instrument using the titanium chemistry generating 700,000 reads, a custom-made perl-script was used for demultiplexing the reads. About 3,500 samples obtained more than 20 reads which had been previously estimated to be sufficient for confident genotyping [190]. This corresponds to 95% of the successful PCR-products which means that almost all vital samples subjected to sequencing were genotyped and this was (and still is?) the largest sample set to be sequenced within one machine run.

In conclusion, the method developed within this project is a good complement when the platform integrated indexing systems are insufficient. Even thou several approaches for sequencing thousands of samples simultaneously have been presented, none of these have actually captured the essence of the problems that occurs when handling really large sample sets. These methods have had the focus on reducing the quantity of unique tags required for a specific sample size and in order to achieve this, sacrificed simplicity.

Since the throughput of today's sequencing instruments is so large, there are resources for targeting larger regions than just one single exon and still investigating thousands of samples together. An adaption of the method described here is presented chapter 3.3 where the double tagging system is combined with hybridization capture and Illumina sequencing.

3.2 Paper II - sorting glowing beads

The idea of this project occurred during the development of the method presented in paper I. After a pilot run on the 454-machine where a pool of 34 individuals was sequenced, an unexpected large proportion of the generated sequencing data contained an unwanted byproduct, less than 100 bases of length. Prior to sequencing, the by-product had appeared as a smear in the

gel electrophoresis analysis but the sample had been purified by gel-cut prior to emPCR and still, the by-product presence in the data was abundant. This is where the idea of a method that would enrich the wanted molecules rather than trying to get rid of the unwanted emerged. Since the PCR-tagging, 454-sequencing and FACS-enriching of DNA carrying beads were already set-up in the group, the idea could be tested relatively quick and the method proved to be useful.

Background

Amplicon sequencing is an effective strategy for deep examination of small genomic regions and the capacity of the next generation sequencing machines allows this to be done in a multitude of samples simultaneously which even further increases the utility of the method. As mentioned in chapter 2.1, PCR may generate unwanted side-products and primer-dimers. If these molecules are not removed in amplicon sequencing experiments prior to library preparation and emPCR, they will steal some of the sequencing capacity. Additionally, if they are shorter than the main product, which is common since they are more likely to occur if they are, they will be more effectively amplified in the emPCR and thereby dominating the sequence data. Applying size selection prior to library preparation is a common way of removing by-products but this requires that the unwanted molecules are of a sufficient different size to be discriminated from the main product and even if the sample looks pure in a gel electrophoresis analysis, it may still contain enough molecules for disturbing the sequencing experiment.

Aim

To make amplicon sequencing more effective, a strategy for enriching emPCR beads carrying amplified target molecules using fluorescence activated cell sorting (FACS) was developed.

Samples

The sample set consisted of a 34 individuals subset of the samples sequenced in paper I. All samples were collected by buccal swabbing and FTA-cards.

Methods

The emPCR beads were labeled using two different fluorophores: *SYBR Green*, that binds all dsDNA and is used for discriminating between naked beads and DNA carrying beads. The other fluorophore, *Alexa647* is coupled to a 50 bp synthetic oligo nucleotide complementary to the central part of the amplicon so that it should not bind primer sequences but still be of sufficient length to tolerate a few mismatches. Based on the fluorophore signal, amplicon carrying beads can be FACS-enrichhed.

To evaluate the utility of enriching target carrying emPCR beads prior to sequencing, three identical sequencing libraries were prepared from the same amplicon pool of 34 individually dual-tagged samples. The amplification target was the same as in paper I: the 2^{nd} exon of the hyper variable DLA-DRB1 gene. One library was attributed to traditional size-selection by gel-cut and the other two were also target enriched using our new approach. All libraries were sequenced in a 1/16 lane in a 454 GS FLX run.

Results and discussion

The data sets from the FACS-enriched sequencing libraries had a nearly three-fold increase of quality reads compared to the traditionally purified control library. The read length of the FACS-enriched libraries was also improved compared to the control library where 69% and 75% of the reads were longer than 200 bp compared to 41%. Since the amplification target is situated within a hyper variable region, single nucleotide substitutions are likely to occur within the probe binding region which also proved to be the case among the 34 individuals included in the experiment. Mismatches in DNA hybridization affects the binding strength so discrimination of individuals non-identical to the DNA probe in the FACS-enrichment process leading to biased data might be a valid hypothesis worth further investigation. Among the 34 samples, up to five mismatches were observed and to investigate if this affected the enrichment efficiency, the reads for each sample was counted within each library and compared between libraries. The largest difference between sample read frequencies observed in the FACS-enriched libraries compared to the control library was 3% but when analyzing the occurrence of probe mismatches among samples with large differences, there was no trend of samples with many mismatches being present in lower fre-

quency in the FACS-enriched libraries. This suggests that the differences observed are likely to be of stochastic nature.

In conclusion, this project resulted in a highly efficient method for minimizing the effect of unwanted by-products in amplicon sequencing experiments. The strategy of using sequence specific fluorescently labeled probes for FACS-enrichment was further investigated in the group and resulted in a method for normalizing pooled MID-libraries [191]. Another utility of this approach is to run a multiplex PCR followed by enrichment of several amplification targets. However, the sequencing costs have declined since this project and today it might be both easier and cheaper to just sequence deeper and filter the by-product reads in the data analysis.

3.3 Paper III - the library

This project was initiated in October 2010 at the Department of Medical Epidemiology and Biostatistics of Karolinska Institutet. At that time there were few third-part library preparation kits on the market so sequencer users had to rely on the expensive reagents provided by the sequencing platform companies. There was also a lot of sample collection going on in the group from tumor tissue and other limited sources and this is where the need for a cheap, efficient and easy method for preparing sequencing libraries compatible with the Illumina sequencing chemistry occurred.

Background

All of today's sequencing platforms needs to incorporate universal sequencing handles in the ends of the molecules to be sequenced to enable clonal amplification and sequencing using universal primers. This process, commonly referred to as library preparation, is discussed in detail in chapter 2.4 but briefly, it involves fragmentation, end-modification and adapter ligation followed by enrichment of correctly ligated molecules by PCR. The effectiveness of this process and the amount of starting material determines the quality of the library and sequentially the quality of the generated data. Hence, if the available amount of input material is limited, a highly effective process is desirable.

While the sequencing costs decreases, the proportional library preparation cost within a sequencing experiment becomes larger which creates a desire for finding cheaper alternatives than whats offered by the sequencing companies. In targeted sequencing experiments, the capture reaction is the most expensive step in the library preparation procedure so performing capture on multiple tagged samples simultaneously has the potential of lowering the total experiment cost significantly.

Aim

This project aimed to reduce costs, hands-on time and DNA consumption in shotgun whole genome sequencing experiments as well as targeted sequencing experiments. The strategy was to develop an effective, cheap and simple protocol for library preparation by eliminating as many cleanup steps as possible and optimizing the reaction conditions so that the different enzymatic steps are conducted in the same buffer by only varying the temperature. Multiplex target capture was also enabled by combining the library preparation protocol and target capture by hybridization.

Samples

The DNA samples used in this study were obtained within an epidemiological study called *Cancer of the Prostate in Sweden* (CAPS) which is a population-based case-control study aimed to investigate genetic and dietary risk factors associated with prostate cancer. Cases were identified from registers between 2001 and 2003. CAPS consists of a little more than 3,000 cases and 2,000 controls which have been included in several GWAS [192, 193] and hence, SNP-chip data is available.

Methods

In the standard library preparation procedure provided by Illumina, three column or bead based clean-up steps is required between fragmentation and PCR enrichment. According to *Qiagen*, a manufacturer of column based clean-up systems, the yield of one clean-up is 60-80% and sequentially, the yield from three clean-up steps is 20-50%. In Illumina's protocol, the adenylation is carried out by the DNA polymerase *Klenow exo*$^-$ which catalyzes the addition of a nucleotide to the 3'end of dsDNA fragments. Since this enzyme

does not distinguish between different nucleotides, only 1/16 of the fragments will be adenylated in both ends if all four nucleotides are present in the reaction. Since they needs to be present in the end-repair reaction, a purification step has to be included prior to adenylation to ensure its effectiveness. By replacing Klenow exo$^-$ with an enzyme that specifically incorporates dATP's even thou all four nucleotides are present, this clean-up step can be avoided. As it happens to be, the most widely used enzyme for PCR namely the *Taq DNA Polymerase* exhibits this property. The name *Taq* is short for *Thermus aquaticus*, which is the thermophilic bacteria the enzyme was isolated from, discovered in the hot springs of the Yellowstone National Park [194]. Taq DNA polymerase has a very low activity at ambient temperatures and needs to be heated to 72°C in order to function optimal. This means that all enzymes involved in end-repair, phosphorylations and adenylation (T4 DNA polymerase, T4 PNK and Taq DNA polymerase respectively) can be added to the same reaction and by altering the reaction temperature, the end-repair and phosphorylation can take place at 25°C and then the adenylation at 72°C whiles the two former enzymes are inactivated by the heat.

In Illumina's library preparation protocol, the reaction buffer needs to be exchanged and the adenylation enzyme removed prior to the ligation reaction but by conducting all reactions in ligase buffer and since the former enzymes are either heat inactivated or inactive at ambient temperatures, this purification step can be skipped. After Y-adapter ligation however, excess adapter constructs needs to be removed prior to enrichment by PCR or they will form extensive amounts of primer-dimer by-products.

In the PCR enrichment step, sample specific barcodes are introduced enabling multiplex sequencing but also multiplex target capture for exome sequencing experiments. To enable multiplex capture of even greater multitudes of samples using dual-tagging, the protocol can be modified by ligating another barcode adapter prior to Y-adapter ligation. This allowed for capturing a 500 kb region in 96 samples simultaneously.

For studying the outcome of different alterations in the library preparation procedure, automated capillary electrophoresis and qPCR was used for assess size distributions and amount of molecules available for amplification using universal primers respectively. The library complexity was assessed by

measuring the PCR-duplicate prevalence in the sequencing data using computational tools like `rmdup` from the alignment manipulating utility collection `SAMtools` [195] and `MarkDuplicates` from `Picard` [196].

Results and discussion

Several factors were varied for improving the efficiency of the procedure. A lowered concentration of Taq DNA polymerase was tested since even though its activity is low at ambient temperature it is still bound to the ends of the DNA fragments and hence, hinders the DNA ligase from binding. Prolonged ligation reaction duration and altered temperatures for phosphorylation and end-repair was also tested. What showed to have an effect was the lowered Taq concentration and the prolonged ligation duration.

Multiplex exome capture was carried out on 2, 4 and 8 samples simultaneously from 1 and 0.1 μg DNA and each capture library was sequenced on $1/8-1/6$ lane on a Illumina HiSeq 2000 machine resulting in an overall mean coverage of 42×. Since the same amount of bait molecules is used for several samples in a multiplex capture, the number of molecules per sample becomes less than in a single-pled capture. This might result in molecules not identical to the reference allele used in the bait design becomes less prone to be captured because of competitive hybridization. To investigate if such bias was introduced in the multiplex captures, the heterozygote variants called from the exome data was compared to SNP-chip data resulting in a concordance of 99.4% (requiring $> 15\times$ coverage). Neither degree of multiplicity nor input DNA amount could be associated to level of concordance suggesting that the underlying mechanism behind variation in captured allele frequencies is of stochastic nature. Further investigations also showed that the multiplex capture did not affect classical parameters used for measure the quality of capture experiments such as GC-content bias, insert-size and fold base 80 penalty. The same comparisons were made for the 96-plex capture as well with 99.8% concordance between heterozygote variant calls and SNP-chip data (requiring $> 15\times$ coverage) and no observed effect on the other parameters.

In conclusion, this project demonstrated an improved efficiency and reduced costs for preparing libraries suitable for both whole-genome and targeted

sequencing experiments using next generation sequencing platforms. The amount of DNA required for building a sequence library was reduced compared to the manufacturer's protocol but there are still other methods more suitable for ultra-low DNA amounts such as *Nextera* (Illumina) [173] and *ThruPLEX* (Rubicon Genomics). Still, the investigations of different factors that affects efficiency and the multiplex capture procedure is useful information for the sequencing community and several researchers who have read the paper have contacted us requesting a detailed protocol.

3.4 Paper IV - DNA in circulation

This have been the main project for this thesis and is also the project where most time have been spent during the last two and a half years. Initiated in september 2010 when circulating tumor DNA had recently been shown to be a promising biomarker in various forms of cancer [80, 81, 93, 94] it was exciting working with something that could potentially change the way cancer is handled. Several approaches for detecting circulating tumor DNA have been tested on a multitude of samples generating tons of results of which only a minor part ended up in the resulting manuscript. Hence, this is a good place to present an extended edition of this project.

Background

There is no one questioning the utility of assessing cancer disease progression and treatment outcome from a single (or a series of) blood sample(s). Using free circulating tumor DNA (ctDNA) as a measure of the systemic tumor load have been proposed and investigated by several studies [80, 81, 93, 94]. However, many of these have focused on patients with advanced disease and a large tumor burden having a relatively large proportion tumor DNA. But to be able to use ctDNA for early relapse detection, surgical outcome measurement and early diagnosis, a robust detection of sub-1% levels of ctDNA must be established. This has been achieved using various approaches such as structural rearrangement breakpoint detection [93,94], sequencing tagged amplicons of hot-spot mutations [197] and counting magnetic fluorescently genotype specific labeled particles using flow cytometry [80, 81, 198].

However, the use of exome sequencing for detecting ctDNA have not yet

been fully explored, even though it has been applied to monitor the genetic evolution in response to therapy in metastatic cancers [199]. Since a vast majority of all solid tumors harbors between 50-100 non-synonymous mutations [40], exome sequencing of blood samples could be a universal tool to assess these in a clinical setting. The challenges of applying this method is the small amounts of available free circulating DNA and the low levels of ctDNA.

Aim

This project aimed to investigate if it is possible to perform captured sequencing form small amounts of starting material, to assess the lower limit of ctDNA detection, to benchmark this against a PCR-based method and to investigate the presence of ctDNA in a small set of clinical samples.

Samples

The samples were selected from the STHLM2 study on the basis of the mutation status of TP53 and recurrent disease. STHLM2 is a cohort study of 25,000 men with he aim of identifying and validate biomarkers of diagnostic and prognostic value for prostate cancer. The collected material consists of questionnaire data and blood samples. In total, three patients were selected that had a mutated TP53, PSA relapse and late stage plasma and urine samples available. Also, breast tumor tissue and blood samples was collected from eight women who underwent surgery for primary breast cancer at Stockholm South General Hospital (KARMAop-study).

Methods

Several approaches for ctDNA detection were investigated during this project. Firstly, the use of structural rearrangement breakpoint detection using PCR-based approaches was explored. To define the structural rearrangement breakpoints within a tumor, low-pass whole-genome sequencing of tumor tissue and germline DNA was applied. We focused on large deletions (>1kb) and the computational approach for finding these is to identify pair-end reads where the insert size is abnormally large. If the mean insert size is increased in a sequencing experiment, a higher physical coverage of the genome will be obtained at a given sequence coverage and hence, increasing the probability

of identifying breakpoints. However, there are some restrictions of how long fragments that can possible be sequenced using the Illumina platform. If the fragments are too long, each cluster will occupy a larger area and hence, the risk of overlapping clusters increases. This can be compensated by providing less molecules to the cluster generation but consequently, this will decrease the number of clusters and obtained reads. If the gain in physical coverage comes with the cost of sequence depth, nothing is won so finding a balance between longer libraries and conserved cluster density is of the essence.

Once some candidate deletions have been identified from the sequencing data, these have to be validated before they can be used on precious plasma samples. This was done by running PCR over the breakpoint using tumor and germline DNA and those breakpoints generating a product in the tumor sample only was considered useful. To use these breakpoints for ctDNA detection, some kind of method that quantifies their presence has to be applied. We used a TaqMan PCR assay [200, 201] combined with digital droplet PCR (ddPCR). We applied this method on five breast cancer patients.

Since the process of developing and evaluate a specific assay for detecting breakpoints for each tumor became very tedious, we also investigated the use of exome sequencing for detecting tumor specific SNV's. In a clinical setting, it is preferable if the same experimental set-up can be used for all patients which is achievable with a sequencing based approach and by targeting a sufficient large part of the genome to ensure enough variation to be captured within most tumors. There are two major challenges for this approach: one is to generate enough sequencing library molecules to be subjected to sequence capture from the relatively small amounts of DNA present in a plasma sample. The second is to be able to detect tumor variants at ultra-low frequencies and to distinguish them from sequencing error induced background signals. To evaluate these limitations, we conducted exome sequencing experiments using various library preparation methods and input DNA amounts. We then used the sequence data *in silico* to estimate the level of background signals and compare them to the tumor signal from samples of various tumor amount by statistical testing. Finally, we applied exome sequencing on a set of clinical samples.

To benchmark the exome sequencing approach, we also amplified tumor spe-

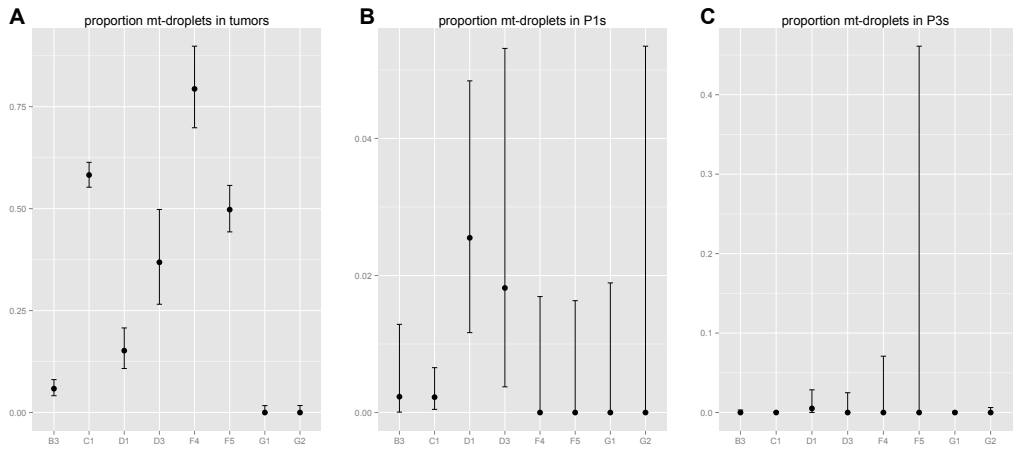

Figure 3.1 Results from the analysis of ctDNA in five breast cancer patients (B-G) using ddPCR. **A:** proportion tumor allele detected in DNA extracted from tumor tissue. **B:** tumor allele in plasma collected just before surgery (P1). **C:** tumor allele in plasma collected after surgery (P3).

cific SNV's located in the gene TP53 and investigated the amplicons using ultra-deep sequencing. This is a robust and easy method and to evaluate the limit of detection, we constructed a synthetic ctDNA dilution series and processed all dilution steps in five replicates. To investigate the impact of template DNA length on the ability of detection, we mirrored the dilution series using full length genomic DNA.

Results and discussion

TaqMan assays for detecting ctDNA was designed for five breast cancer patients and evaluated using conventional qPCR. Using ddPCR, ctDNA could be detected in two out of five breast cancer patients (C and D) in blood samples taken just before surgery (see figure 3.1). Out of the five patients, these were the two that harbored the largest tumors. In sample D1P3, trace amounts of ctDNA could be detected at a non-significant level. Since we used a demonstration ddPCR-instrument, kindly provided by Bio-Rad, we only had one day of use and therefore, we did not had the time to optimize the PCR conditions for each assay. As seen in figure 3.1 A, the assays G1 and G2 did not work at all and there is great variance in the detected tumor allele proportion using different assays within the same patient. This need

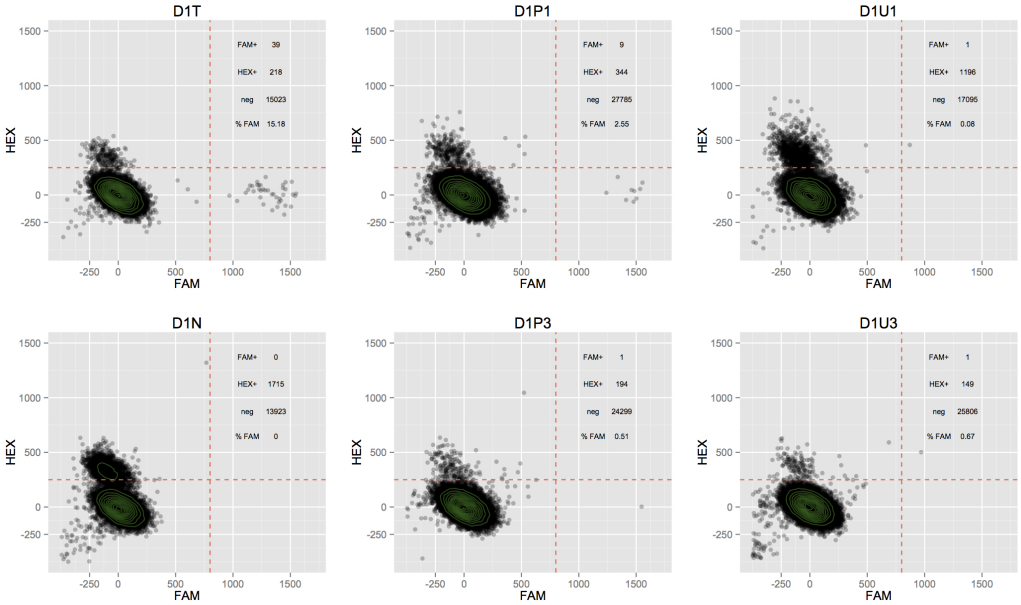

Figure 3.2 Example of a allele discrimination plot from assay D1 (N=normal, T=tumor, P=plasma, U=urine). On the y-axis is the signal from the wild-type probe and on the x-axis is the signal from the mutant probe.

of individual assay optimization makes this approach less suitable for clinical implementation. Digital droplet PCR provided a high resolution analysis of rare alleles and since two different fluorophores were used, one for the deletion breakpoint and one for the wild type of the forward primer side of the deletion, a two-dimensional map of the allele frequency for each droplet can be constructed for each sample (see figure 3.2) However, there is a lot of evaluation and optimization behind each assay which makes this method less suitable in a clinical setting and further more, the cost of each assay is high. Because of this, we decided to exclude this approach from the manuscript.

To evaluate the utility of captured sequencing from minute amounts of starting material, we let two different library preparation methods, both of which is dedicated to prepare sequencing libraries from small amounts of DNA, duel against each other by performing sequence capture and sequencing from 1 ng and 10 ng respectively. The outcome was the complexity of the capture libraries (the concept of library complexity is further discussed in chapter 2.4) measured as the average sequence coverage after removing PCR duplicates.

The two combatants were the ThruPLEX kit (Rubicon Genomics) and the Mondrian system (NuGEN), both of which should be capable of generating sequencing libraries from nano grams of DNA according to their respective manufacturer. The average coverage using 1 ng and 10 ng of starting DNA was 1.7× and 2.9× for the Mondrian system and 11.7× and 60.8× using ThruPLEX. Thus, the superior winner was ThruPLEX which in addition, had a very simple laboratory procedure.

The data generated in the comparison experiment was also used for determine the sensitivity of ctDNA detection using captured sequencing. This was done by *in silco* estimation of background signal and generation of simulated datasets of various tumor allele content. This resulted in an estimated detection level of 0.5% with 95% sensitivity.

Sequencing an amplified tumor specific SNV for ctDNA detection was also evaluated using a series of synthetic plasma DNA samples of various amount of ctDNA. The different degrees of tumor allele content was processed in 5× replicates and in the samples of lowest tumor allele content (aimed to be 0.025% and estimated to 0.03%), significant detection of tumor allele could be achieved in 1 out of 3 replicates (the other two replicates experimentally failed). The nature of PCR demands that the region to be amplified must be continuously represented within a molecule to enable amplification and hence, a fragmented sample must contain less molecules available for amplification than a non-fragmented sample. To test this hypothesis, a similar series of tumor DNA diluted in germline DNA using non-fragmented sources was made. The result from this was that the sample of lowest tumor content had a mean estimate of 0.09% (again supposed to be 0.025%) and two out of four replicates obtained significant detection. The estimate's difference from the estimate of the corresponding fragmented DNA sample might be from pipetting errors as well as stochastic effects. No difference in the coefficient of variance between replicates was observed either. Nevertheless, all replicates of the second lowest tumor DNA content (supposed to be 0.05%) obtained significant detection of tumor allele with mean estimates of 0.25% and 0.17% for the non-fragmented and fragmented DNA respectively.

This PCR-based approach was applied on a series of clinical plasma and urine samples from three prostate cancer patients taken at various time points. Sur-

prisingly, they showed undetectable or very low levels of ctDNA even though they were taken from patients having PSA relapse and a muted TP53 allele. In the plasma taken from the patient having the highest PSA-value (114 ng/ml), 0.13% tumor allele was detected which stands in contrast to previously published work where patients with PSA levels between 17.4 ng/ml and 808 ng/ml had ctDNA levels of 30-50% [202]. However, another study has recently reported about severe colon cancer cases having undetectable or very low ctDNA levels [203]. This suggests that there may be tumor phenotypes that are simply not so very prone to share their genetic material to their surroundings and perhaps, these are more common within prostates? Another explanation to the low ctDNA levels observed here might be that the blood samples were not handled in an optimal way for ctDNA analysis. Ideally, the blood should be spun and frozen within a few hours after it has been taken but for some of the samples analyzed here, the needle-to-freezer time was up to three days. During this time, blood cells might die and leak DNA to the sample and also, nucleases may be active (in spite of the high EDTA concentration) and degrading the free DNA. If these two processes occurs simultaneously, the ctDNA will be heavily diluted in germline DNA whiles the total concentration of free DNA remains somewhat unaffected.

We also applied exome sequencing to the same clinical samples but since their ctDNA levels was below our *in silico* estimated level of detection, there was no ctDNA detected.

To summarize, we showed that the ability of detecting low proportions of ctDNA using exome sequencing is currently limited by the noise levels using Illumina sequencing. Nevertheless, we did show that exome sequencing can be achieved from small amounts of starting material. Since information about the mutational status of driver genes and genetic indications of therapy resistance also is obtained, it might be of good use for monitoring treatment outcome in severe cases having high ctDNA levels. The PCR-based methods proved to be useful for detection of lower levels of ctDNA. Sequencing an amplified SNV proved to be both simpler and cheaper than running ddPCR over a breakpoint. It also indicated to have a lower limit of detection, even though a fair comparison never was done. Sequencing amplified SNV's is attractive because of its ease of use and is also available for more labs, since having a massive sequencer is more common than having a ddPCR machine.

3.5 Future perspectives

The papers presented in this thesis have all enabled more efficient investigations of DNA using sequencing. This is a filed that currently develops at a very high pace and it is not far to conclude that technical achievements rapidly becomes out of date. Nevertheless, the findings presented in this thesis all aimed to make the most use of the massive throughput and enabling new applications rather than to improve it. The tagging procedure in paper I will be even more applicable in the future since it not only enables the analysis of thousands of samples simultaneously, but also facilitates a logistic strategy to handle all these samples in the laboratory. This will be of importance when even larger sample sets are to be processed.

The cost per sequenced base has decreased and will probably continue to fall (maybe at a slightly lower pace) during the years to come which makes the tolerance to data contamination and uneven read distributions across samples larger. This might make the method of sorting DNA carrying beads, presented in paper II, redundant but there is still some potential if larger sets of targeted genes can be enriched from more complex sequencing libraries. The library preparation and target enrichment procedure presented in paper III has already been requested by several researches that have read the article and perhaps the time when these findings are of most importance is now? If the future holds library-free sequencing methods at extremely high throughput, a method for preparing sequencing libraries and enriching genomic regions might not be essential.

The utilization of ctDNA for handling cancer is in its infancy and has not yet been established in clinical practice. Currently, various methods are being evaluated while trying to understand the biological mechanisms. The investigation of the limitations using exome sequencing for ctDNA applications is therefore of great value. Knowledge of a weakness must be gained before it can be turned into a strength and there is a high probability that assessment of ctDNA using sequencing will be clinically implemented in the future.

During the last years, we have witnessed an explosion in sequencing technologies. Be aware, because the biological knowledge that follows is about to burst.

Populärvetenskaplig sammanfattning

Alla celler innehåller en mängd olika proteiner som utför olika uppgifter, till exempel energiproduktion, kommunikation, underhåll, transporter och mekanisk hållfasthet. För att en cell skall kunna producera dessa proteiner behöver den information om hur de skall sammansättas, en ritning. Denna information finns lagrad i en molekyl som kallas DNA som är en lång kedja, uppbyggd av fyra olika bitar som kallas baser (A, C, G och T). På samma sätt som bokstäverna i den här texten bildar ord bestäms informationen lagrad i DNA-molekylen av ordningen, eller sekvensen av baserna. Ritningen för ett visst protein kallas *gen* och samlingen av alla ritningar hos en organism kallas *genom*. Människans genom består av cirka tre miljarder (3×10^9) baser som kodar för 20,687 gener och skulle man skriva ut alla dessa baser i en rad med 1 mm breda bokstäver skulle den raden vara lika lång som resan från Stockholm till Paris tur och retur (3,137 km). I denna skala skulle det finnas en gen var 300:e meter och varje gen skulle vara i snitt 27 meter lång. När man undersöker DNA brukar man vilja ta reda på i vilken ordning baserna sitter i och denna process kallas för att *sekvensera*.

Cancer är en sjukdom som uppkommer för att baser ibland blir utbytta mot andra baser, man säger att det sker en *mutation*. Detta kan hända på grund av att cellens kopieringsmaskineri inte är helt perfekt när en cell kopierar hela sitt genom inför en celldelning. En mutation kan vara både skadlig, gynnsam eller inte alls påverka en cell och detta är en viktig mekanism som ligger bakom det som kallas för evolution. En mutation kan också uppstå när en cell utsätts för något i dess miljö som påverkar DNA-molekylen på ett skadligt sätt, som till exempel ultraviolett strålning eller tobaksrök. En cancercell har samlat på sig mutationer som gett den en överlevnadsfördel gentemot andra

celler i dess närmiljö och kan därför växa snabbare och mer okontrollerat än sina grannar. Som tur är tar detta väldigt lång tid och det är därför cancer ofta uppkommer sent i livet.

I Sverige är prostatacancer den vanligaste formen av cancer hos män och varje år får cirka 9,500 personer diagnosen, vilket motsvarar ungefär en tredjedel av alla manliga cancerfall. Prostatan är en körtel lika stor som en valnöt som sitter under urinblåsan och omsluter urinröret. Det lömska med prostatacancer är att sjukdomen är i princip symtomfri i tidigt skede och på grund av prostatans lokalisering är det även svårt att upptäcka en tumör med fingrarna. Prostatan utsöndrar ett protein som kallas PSA och genom att mäta halten av detta protein i blodet kan man upptäcka om prostatan växer vilket kan vara ett tecken på en tumör. För att utreda om en patient med ett högt PSA-värde verkligen har prostatacancer tar man en biopsi för att undersöka hur cellerna ser ut och utifrån denna kan en diagnos ställas. Genom att titta på hur mycket cellerna har förändrats, hur stor tumören är samt hur högt PSA-värde patienten har kan man avgöra om sjukdomen är aggressiv. För behandling av prostatacancer skiljer man på fall där hela tumören växer inuti prostatan, så kallad *lokaliserad* sjukdom och fall där tumören har spridits till omgivande vävnad eller andra organ, så kallad *metastaserade* sjukdom. Lokaliserad prostatacancer kan behandlas med strålning, kirurgi eller ibland inte alls beroende på patientens ålder, patientens allmäntillstånd samt tumörens aggressivitet. Metastaserade prostatacancer behandlas oftast med läkemedel som blockerar de hormoner som stimulerar tumören till att växa.

I den här avhandlingen har en metod för att följa en tumörs utveckling med hjälp av blodprover undersökts. Cellerna i en cancertumör är mer biokemiskt aktiva än normala celler och en tumör utsöndrar därför också mer material till sin omgivning än vad frisk vävnad gör. Bland annat utsöndras DNA och eftersom en tumör bär på mutationer är detta DNA annorlunda i jämförelse med resten av kroppens cellers DNA. Genom att sekvensera tumörvävnad kan man ta reda på vilka baser som är muterade i en specifik tumör. Dessa kan man sedan leta efter i det fritt cirkulerande DNA som finns i blodet och på så sätt får man ett mått på hur stor och hur aktiv en tumör är. Det svåra är att det även finns fritt cirkulerande DNA från friska celler i blodet och halten av tumör-DNA är ibland väldigt låg (mindre än 1%). Det har

vi löst genom att sekvensera de muterade positionerna väldigt noga för att på så sätt få ihop tillräckligt många observationer för att statistikt kunna säkerställa närvaro av tumör. Det som begränsar metoden är att den maskin som utför själva sekvenseringen ibland gör fel så att det ser ut som om en bas är muterad fastän den i själva verket inte är det. Det gäller alltså att den muterade basen från tumören finnas närvarande i högre utsträckning än sekvenseringsfelen för att kunna identifieras.

Den här metoden innebär att man kan följa hur en tumör svarar på olika behandlingar. Man kan till exempel se att halten tumör-DNA sjunker när patienten får cellgifter och att den stiger om en patient får ett återfall. Hos en patient med hög halt tumör-DNA i blodet kan man även se om en tumör får nya mutationer, som till exempel resistans mot en viss behandling. Ett framtida alternativ till operation kan vara att ta fram en genetisk profil för tumören utifrån en biopsi och sedan följ dess utveckling med hjälp av cirkulerande tumör-DNA. Skulle man se att halten stiger och att tumören fått gener associerade med aggressiv sjukdom muterade kan man besluta om operation. På så sätt skulle många operationer som idag görs i onödan kunna undvikas.

Acknowledgments

Det bakomliggande arbetet till denna bok hade inte varit möjligt att genomföra utan bra handledare, skickliga medarbetare, trevliga kollegor och stöttande familjemedlemmar. Här kommer en massa tack:

Henrik Grönberg, *huvudhandledare*, tack för att du gick mot normen och tog in en tekniskt inriktad doktorand till MEB. Det har varit otroligt berikande för mig att fått se saker ur ett epidemiologiskt perspektiv och fått vara inblandad i klinisk forskning. Man är alltid inspirerad och motiverad efter ett möte med dig och ditt engagemang för teknikutveckling har varit mycket värdefullt.

Daniel Klevebring, *bihandledare*, utan dig hade jag inte fortsatt mina forskarstudier. Tack för att du trodde på mig och gav mig chansen att fullfölja mina forskarstudier. Du är inte bara en entusiastisk och otroligt duktig forskare, du är även en god vän. Tack också för alla nördiga samtal över en frukost på Mellqvist's.

Johan Lindberg, *bihandledare*, det har varit väldigt spännande att jobba med dig och jag är väldigt glad att ha fått lära känna dig. Din skicklighet och extrema koll på den vetenskapliga litteraturen har varit stora inspirationskällor. Det har även varit berikande med alla film-, ekonomi-, kaffe- och familjediskussioner.

Julia, det var verkligen roligt att du började på MEB. Tack för alla luncher, fikastunder och samtal och tack för att du är en god vän.

Tack **Simon** och **Anna** för ett fantastiskt arbetsklimat i vårt lilla hörn på SciLife, det har varit än ära att få jobba med er. Tack också **Gabriela** för att du förgyller vår tillvaro när du kommer och hälsar på.

Tobias och **Marcus**, tack för alla intressanta samtal, kliniska perspektiv och trevliga luncher. Tack också **Fredrik** för trevliga luncher och statistikdiskussioner.
Thank you **Kamila** and **Per** for being great sources of inspiration and for letting me have a piece of KARMA.
Tack **Lars** och **Peter** och alla andra kliniskt verksamma medarbetare på Karolinska Sjukhuset. Tack också till gänget på Södersjukhuset, speciellt **Louise**, **Fuat** och **Eva**.
Tack **Robert** och **Robert** för trevligt rumssällskap på Lidö och i Montreal och tack **Martin** för att du tog hand om mig när jag började på MEB.
Tack alla medarbetare i **prostatagruppen** för att ni är så trevliga och ett extra tack till **Karin** som tog blodprov på mig och Tobias.
Camilla och **Gunilla**, tack för all hjälp med administration.
Tank you **all nice MEB'ers** for contributing to a friendly and stimulating research environment.

Mina forna KTH-handledare **Afshin**, **Peter** och **Joakim**, tack för den tid jag spenderade hos er och för allt jag lärde mig.
Sverker, tack för mycket trevliga luncher och djupa diskussioner. Jag är stolt över att jag nästan fick dig att börja klättra.
Klättergänget på SciLife bestående av **Anders**, **Erik**, **Anna**, **Mattias**, **Måns**, m.fl., tack för trevliga klättersessioner och mumsiga burgare.
Jochen, thanks for your encouragement.
Tack alla i **SciLife's core-verksamhet** som bistått med expertis, infrastruktur och samarbeten.

Tack **mamma och pappa** för er support, uppmuntran och för ert engagemang. Jag hade aldrig kommit såhär långt utan er!
Jag vill även tacka min syster **Åsa**, hennes man **Magnus**, min svärmor **Margareta**, min svärfar **Öivind** och alla andra familjemedlemmar som hejar på mig!

Maja och **Märta**, ni är de viktigaste i mitt liv. Jag trivs så bra med er i vårt lilla hus bland talltopparna. Märta, jag är så glad över att vara pappa till dig. Maja, du är den finaste fru och livskamrat man kan ha. Jag älskar dig!

Bibliography

[1] R. Dahm and F. Miescher. Discovering DNA: Friedrich Miescher and the early years of nucleic acid research. *Hum. Genet.*, 122(6):565–581, Jan 2008.

[2] O. T. Avery, C. M. Macleod, and M. McCarty. Studies on the chemical nature of the substance inducing transformation of pneumococcal types. Induction of transformation by a desoxyribonucleic acid fraction isolated from Pneumococcus type III. *J. Exp. Med.*, 79(2):137–158, Feb 1944.

[3] J. D. Watson and F. H. Crick. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–738, Apr 1953.

[4] F. Crick. Central dogma of molecular biology. *Nature*, 227(5258):561–563, Aug 1970.

[5] University of California Santa Cruz (UCSC) Genome Bioinformatics Group. Hg19 genome size statistics. *website*, (`http://genomewiki.ucsc.edu/index.php/Hg19_Genome_size_statistics`), February 2013.

[6] E. S. Lander et al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, Feb 2001.

[7] J. C. Venter et al. The sequence of the human genome. *Science*, 291(5507):1304–1351, Feb 2001.

[8] N. H. HOROWITZ. Progress in developing chemical concepts of genetic phenomena. *Fed. Proc.*, 15(2):818–822, Jul 1956.

[9] D. M. BONNER. The genetic unit. *Cold Spring Harb. Symp. Quant. Biol.*, 21:163–170, 1956.

[10] S. Ohno. So much "junk" DNA in our genome. *Brookhaven Symp. Biol.*, 23:366–370, 1972.

[11] The ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science*, 306(5696):636–640, Oct 2004.

[12] International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011):931–945, Oct 2004.

[13] The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, Sep 2012.

[14] A. J. Stunkard, T. T. Foch, and Z. Hrubec. A twin study of human obesity. *JAMA*, 256(1):51–54, Jul 1986.

[15] 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073, Oct 2010.

[16] P. D. Stenson, M. Mort, E. V. Ball, K. Howells, A. D. Phillips, N. S. Thomas, and D. N. Cooper. The Human Gene Mutation Database: 2008 update. *Genome Med*, 1(1):13, 2009.

[17] 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65, Nov 2012.

[18] A. S. Wiener. Method of Measuring Linkage in Human Genetics; with Special Reference to Blood Groups. *Genetics*, 17(3):335–350, May 1932.

[19] K. Yoshiura et al. A SNP in the ABCC11 gene is the determinant of human earwax type. *Nat. Genet.*, 38(3):324–330, Mar 2006.

[20] R. A. King, J. Pietsch, J. P. Fryer, S. Savage, M. J. Brott, I. Russell-Eggitt, C. G. Summers, and W. S. Oetting. Tyrosinase gene mutations in oculocutaneous albinism 1 (OCA1): definition of the phenotype. *Hum. Genet.*, 113(6):502–513, Nov 2003.

[21] D. L. Duffy, G. W. Montgomery, W. Chen, Z. Z. Zhao, L. Le, M. R. James, N. K. Hayward, N. G. Martin, and R. A. Sturm. A three-single-nucleotide polymorphism haplotype in intron 1 of OCA2 explains most human eye-color variation. *Am. J. Hum. Genet.*, 80(2):241–252, Feb 2007.

[22] G. M. Cooper, J. A. Johnson, T. Y. Langaee, H. Feng, I. B. Stanaway, U. I. Schwarz, M. D. Ritchie, C. M. Stein, D. M. Roden, J. D. Smith, D. L. Veenstra, A. E. Rettie, and M. J. Rieder. A genome-wide scan for common genetic variants with a large influence on warfarin maintenance dose. *Blood*, 112(4):1022–1027, Aug 2008.

[23] A. K. Daly et al. HLA-B*5701 genotype is a major determinant of drug-induced liver injury due to flucloxacillin. *Nat. Genet.*, 41(7):816–819, Jul 2009.

[24] M. McCormack et al. HLA-A*3101 and carbamazepine-induced hypersensitivity reactions in Europeans. *N. Engl. J. Med.*, 364(12):1134–1143, Mar 2011.

[25] R. Mei, P. C. Galipeau, C. Prass, A. Berno, G. Ghandour, N. Patil, R. K. Wolff, M. S. Chee, B. J. Reid, and D. J. Lockhart. Genome-wide detection of allelic imbalance using human SNPs and high-density DNA arrays. *Genome Res.*, 10(8):1126–1137, Aug 2000.

[26] G. C. Kennedy, H. Matsuzaki, S. Dong, W. M. Liu, J. Huang, G. Liu, X. Su, M. Cao, W. Chen, J. Zhang, W. Liu, G. Yang, X. Di, T. Ryder, Z. He, U. Surti, M. S. Phillips, M. T. Boyce-Jacino, S. P. Fodor, and K. W. Jones. Large-scale genotyping of complex DNA. *Nat. Biotechnol.*, 21(10):1233–1237, Oct 2003.

[27] K. L. Gunderson, F. J. Steemers, G. Lee, L. G. Mendoza, and M. S. Chee. A genome-wide scalable SNP genotyping assay using microarray technology. *Nat. Genet.*, 37(5):549–554, May 2005.

[28] Inc. Illumina. HumanOmni5-Quad BeadChip. *Data Sheet: DNA Analysis*, Pub. No. 370-2011-007, October 2011.

[29] J. H. Barrett et al. Genome-wide association study identifies three new melanoma susceptibility loci. *Nat. Genet.*, 43(11):1108–1113, Nov 2011.

[30] R. Karlsson, M. Aly, M. Clements, L. Zheng, J. Adolfsson, J. Xu, H. Gronberg, and F. Wiklund. A Population-based Assessment of Germline HOXB13 G84E Mutation and Prostate Cancer Risk. *Eur. Urol.*, Jul 2012.

[31] J. M. Scharf et al. Genome-wide association study of Tourette's syndrome. *Mol. Psychiatry*, Aug 2012.

[32] T. Freilinger et al. Genome-wide association analysis identifies susceptibility loci for migraine without aura. *Nat. Genet.*, 44(7):777–782, Jul 2012.

[33] P. Sulem et al. Genetic determinants of hair, eye and skin pigmentation in Europeans. *Nat. Genet.*, 39(12):1443–1452, Dec 2007.

[34] B. Kerem, J. M. Rommens, J. A. Buchanan, D. Markiewicz, T. K. Cox, A. Chakravarti, M. Buchwald, and L. C. Tsui. Identification of the cystic fibrosis gene: genetic analysis. *Science*, 245(4922):1073–1080, Sep 1989.

[35] W. S. Bush and J. H. Moore. Chapter 11: Genome-wide association studies. *PLoS Comput. Biol.*, 8(12):e1002822, Dec 2012.

[36] D. E. Reich and E. S. Lander. On the allelic spectrum of human disease. *Trends Genet.*, 17(9):502–510, Sep 2001.

[37] F. Dudbridge and A. Gusnanto. Estimation of significance thresholds for genomewide association scans. *Genet. Epidemiol.*, 32(3):227–234, Apr 2008.

[38] Hindorff L.A., MacArthur J., Morales J., Junkins H.A., Hall P.N., Klemm A.K., and Manolio TA. A Catalog of Published Genome-Wide Association Studies. *website*, (`http://www.genome.gov/gwastudies/`), Accessed April 2013.

[39] B. Alberts, D. Bray, K. Hopkin, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P Walter. *Essential cell biology.* Garland Science, second edition, 2004.

[40] B. Vogelstein, N. Papadopoulos, V. E. Velculescu, S. Zhou, L. A. Diaz, and K. W. Kinzler. Cancer genome landscapes. *Science*, 339(6127):1546–1558, Mar 2013.

[41] R. Govindan, L. Ding, M. Griffith, J. Subramanian, N. D. Dees, K. L. Kanchi, C. A. Maher, R. Fulton, L. Fulton, J. Wallis, K. Chen, J. Walker, S. McDonald, R. Bose, D. Ornitz, D. Xiong, M. You, D. J. Dooling, M. Watson, E. R. Mardis, and R. K. Wilson. Genomic landscape of non-small cell lung cancer in smokers and never-smokers. *Cell*, 150(6):1121–1134, Sep 2012.

[42] C. M. Croce. Oncogenes and cancer. *N. Engl. J. Med.*, 358(5):502–511, Jan 2008.

[43] M. Vazquez, V. de la Torre, and A. Valencia. Chapter 14: Cancer genome analysis. *PLoS Comput. Biol.*, 8(12):e1002824, Dec 2012.

[44] National Cancer Institute (NCI). Dictionary of cancer terms. *website*, (`www.cancer.gov/dictionary`), Mars 2013.

[45] J. A. Ludwig and J. N. Weinstein. Biomarkers in cancer staging, prognosis and treatment selection. *Nat. Rev. Cancer*, 5(11):845–856, Nov 2005.

[46] N. F. Boyd, G. A. Lockwood, J. W. Byng, D. L. Tritchler, and M. J. Yaffe. Mammographic densities and breast cancer risk. *Cancer Epidemiol. Biomarkers Prev.*, 7(12):1133–1144, Dec 1998.

[47] L. M. McShane, D. G. Altman, W. Sauerbrei, S. E. Taube, M. Gion, and G. M. Clark. REporting recommendations for tumor MARKer prognostic studies (REMARK). *Breast Cancer Res. Treat.*, 100(2):229–235, Nov 2006.

[48] E. P. Diamandis. The failure of protein cancer biomarkers to reach the clinic: why, and what can be done to address the problem? *BMC Med*, 10:87, 2012.

[49] The National Board of Health and Welfare, Sweden. Cancer incidience in Sweden 2011. Dec 2012.

[50] H. Gronberg. Prostate cancer epidemiology. *Lancet*, 361(9360):859–864, Mar 2003.

[51] G. Engholm, J. Ferlay, N. Christensen, F. Bray, M. L. Gjerstorff, A. Klint, J. E. K?tlum, E. Olafsdottir, E. Pukkala, and H. H. Storm. NORDCAN–a Nordic tool for cancer information, planning, quality control and research. *Acta Oncol*, 49(5):725–736, Jun 2010.

[52] K. M. Wilson, E. L. Giovannucci, and L. A. Mucci. Lifestyle and dietary factors in the prevention of lethal prostate cancer. *Asian J. Androl.*, 14(3):365–374, May 2012.

[53] A. S. Whittemore, L. N. Kolonel, A. H. Wu, E. M. John, R. P. Gallagher, G. R. Howe, J. D. Burch, J. Hankin, D. M. Dreon, and D. W. West. Prostate cancer in relation to diet, physical activity, and body size in blacks, whites, and Asians in the United States and Canada. *J. Natl. Cancer Inst.*, 87(9):652–661, May 1995.

[54] J. M. Chan, M. J. Stampfer, J. Ma, P. H. Gann, J. M. Gaziano, and E. L. Giovannucci. Dairy products, calcium, and prostate cancer risk in the Physicians' Health Study. *Am. J. Clin. Nutr.*, 74(4):549–554, Oct 2001.

[55] L. E. Johns and R. S. Houlston. A systematic review and meta-analysis of familial prostate cancer risk. *BJU Int.*, 91(9):789–794, Jun 2003.

[56] S. Madersbacher, A. Alcaraz, M. Emberton, P. Hammerer, A. Ponholzer, F. H. Schroder, and A. Tubaro. The influence of family history on prostate cancer risk: implications for clinical management. *BJU Int.*, 107(5):716–721, Mar 2011.

[57] P. Lichtenstein, N. V. Holm, P. K. Verkasalo, A. Iliadou, J. Kaprio, M. Koskenvuo, E. Pukkala, A. Skytthe, and K. Hemminki. Environmental and heritable factors in the causation of cancer–analyses of cohorts of

twins from Sweden, Denmark, and Finland. *N. Engl. J. Med.*, 343(2):78–85, Jul 2000.

[58] C. L. Goh, F. R. Schumacher, D. Easton, K. Muir, B. Henderson, Z. Kote-Jarai, and R. A. Eeles. Genetic variants associated with predisposition to prostate cancer and potential clinical implications. *J. Intern. Med.*, 271(4):353–365, Apr 2012.

[59] R. A. Eeles et al. Identification of 23 new prostate cancer susceptibility loci using the iCOGS custom genotyping array. *Nat. Genet.*, 45(4):385–391, Apr 2013.

[60] C. M. Ewing, A. M. Ray, E. M. Lange, K. A. Zuhlke, C. M. Robbins, W. D. Tembe, K. E. Wiley, S. D. Isaacs, D. Johng, Y. Wang, C. Bizon, G. Yan, M. Gielzak, A. W. Partin, V. Shanmugam, T. Izatt, S. Sinari, D. W. Craig, S. L. Zheng, P. C. Walsh, J. E. Montie, J. Xu, J. D. Carpten, W. B. Isaacs, and K. A. Cooney. Germline mutations in HOXB13 and prostate-cancer risk. *N. Engl. J. Med.*, 366(2):141–149, Jan 2012.

[61] J. Xu et al. HOXB13 is a susceptibility gene for prostate cancer: results from the International Consortium for Prostate Cancer Genetics (ICPCG). *Hum. Genet.*, 132(1):5–14, Jan 2013.

[62] The National Board of Health and Welfare, Sweden. Cancer i siffror 2009. 2009.

[63] K. H. Leissner and L. E. Tisell. The weight of the human prostate. *Scand. J. Urol. Nephrol.*, 13(2):137–142, 1979.

[64] C. Huggins, W. W. Scott, and J. H. Heinen. Chemical composition of human semen and of the secretions of the prostate and seminal vesicles. *Am. J. Physiol.*, 136:467–473, 1942.

[65] K. M. Verhamme et al. Incidence and prevalence of lower urinary tract symptoms suggestive of benign prostatic hyperplasia in primary care–the Triumph project. *Eur. Urol.*, 42(4):323–328, Oct 2002.

[66] M. Yin, S. Bastacky, U. Chandran, M. J. Becich, and R. Dhir. Prevalence of incidental prostate cancer in the general population: a study of healthy organ donors. *J. Urol.*, 179(3):892–895, Mar 2008.

[67] F. H. Schroder et al. Screening and prostate-cancer mortality in a randomized European study. *N. Engl. J. Med.*, 360(13):1320–1328, Mar 2009.

[68] F. H. Schroder. Landmarks in prostate cancer screening. *BJU Int.*, 110 Suppl 1:3–7, Oct 2012.

[69] C. S. Killian, N. Yang, L. J. Emrich, F. P. Vargas, M. Kuriyama, M. C. Wang, N. H. Slack, L. D. Papsidero, G. P. Murphy, and T. M. Chu. Prognostic importance of prostate-specific antigen for monitoring patients with stages B2 to D1 prostate cancer. *Cancer Res.*, 45(2):886–891, Feb 1985.

[70] S. P. Balk, Y. J. Ko, and G. J. Bubley. Biology of prostate-specific antigen. *J. Clin. Oncol.*, 21(2):383–391, Jan 2003.

[71] M. Barry and C. Roehrborn. Management of benign prostatic hyperplasia. *Annu. Rev. Med.*, 48:177–189, 1997.

[72] R. Etzioni, D. F. Penson, J. M. Legler, D. di Tommaso, R. Boer, P. H. Gann, and E. J. Feuer. Overdiagnosis due to prostate-specific antigen screening: lessons from U.S. prostate cancer incidence trends. *J. Natl. Cancer Inst.*, 94(13):981–990, Jul 2002.

[73] H. G. Welch and P. C. Albertsen. Prostate cancer diagnosis and treatment after the introduction of prostate-specific antigen screening: 1986-2005. *J. Natl. Cancer Inst.*, 101(19):1325–1329, Oct 2009.

[74] Jonas Hugosson, Sigrid Carlsson, Gunnar Aus, Svante Bergdahl, Ali Khatami, Pär Lodding, Carl-Gustaf Pihl, Johan Stranne, Erik Holmberg, and Hans Lilja. Mortality results from the göteborg randomised population-based prostate-cancer screening trial. *The lancet oncology*, 11(8):725–732, 2010.

[75] D. F. Gleason. Classification of prostatic carcinomas. *Cancer Chemother Rep*, 50(3):125–128, Mar 1966.

[76] T. J. Sebo, B. J. Bock, J. C. Cheville, C. Lohse, P. Wollan, and H. Zincke. The percent of cores positive for cancer in prostate needle biopsy specimens is strongly predictive of tumor stage and volume at radical prostatectomy. *J. Urol.*, 163(1):174–178, Jan 2000.

[77] J. I. Epstein. An update of the Gleason grading system. *J. Urol.*, 183(2):433–440, Feb 2010.

[78] The National Board of Health and Welfare, Sweden. Nationella riktlinjer för bröst-, prostata-, tjocktarms- och ändtarmscancervård 2013. Mars.

[79] P. Mandel and P. Metais. Les acides nucleiques plasma sanguin chez l'homme. *Acad. Sci. Paris*, 142:241–243, 1948.

[80] F. Diehl, M. Li, D. Dressman, Y. He, D. Shen, S. Szabo, L. A. Diaz, S. N. Goodman, K. A. David, H. Juhl, K. W. Kinzler, and B. Vogelstein. Detection and quantification of mutations in the plasma of patients with colorectal tumors. *Proc. Natl. Acad. Sci. U.S.A.*, 102(45):16368–16373, Nov 2005.

[81] F. Diehl, K. Schmidt, M. A. Choti, K. Romans, S. Goodman, M. Li, K. Thornton, N. Agrawal, L. Sokoll, S. A. Szabo, K. W. Kinzler, B. Vogelstein, and L. A. Diaz. Circulating mutant DNA to assess tumor dynamics. *Nat. Med.*, 14(9):985–990, Sep 2008.

[82] G. D. Sorenson, D. M. Pribish, F. H. Valone, V. A. Memoli, D. J. Bzik, and S. L. Yao. Soluble normal and mutated DNA sequences from single-copy genes in human blood. *Cancer Epidemiol. Biomarkers Prev.*, 3(1):67–71, 1994.

[83] V. Vasioukhin, P. Anker, P. Maurice, J. Lyautey, C. Lederrey, and M. Stroun. Point mutations of the N-ras gene in the blood plasma DNA of patients with myelodysplastic syndrome or acute myelogenous leukaemia. *Br. J. Haematol.*, 86(4):774–779, Apr 1994.

[84] S. A. Leon, B. Shapiro, D. M. Sklaroff, and M. J. Yaros. Free DNA in the serum of cancer patients and the effect of therapy. *Cancer Res.*, 37(3):646–650, Mar 1977.

[85] M. Fleischhacker and B. Schmidt. Circulating nucleic acids (CNAs) and cancer–a survey. *Biochim. Biophys. Acta*, 1775(1):181–232, Jan 2007.

[86] E. Gormally et al. Amount of DNA in plasma and cancer risk: a prospective study. *Int. J. Cancer*, 111(5):746–749, Sep 2004.

[87] Y. M. Lo, N. Corbetta, P. F. Chamberlain, V. Rai, I. L. Sargent, C. W. Redman, and J. S. Wainscoat. Presence of fetal DNA in maternal plasma and serum. *Lancet*, 350(9076):485–487, Aug 1997.

[88] Y. M. Lo, M. S. Tein, T. K. Lau, C. J. Haines, T. N. Leung, P. M. Poon, J. S. Wainscoat, P. J. Johnson, A. M. Chang, and N. M. Hjelm. Quantitative analysis of fetal DNA in maternal plasma and serum: implications for noninvasive prenatal diagnosis. *Am. J. Hum. Genet.*, 62(4):768–775, Apr 1998.

[89] Y. M. Lo, J. Zhang, T. N. Leung, T. K. Lau, A. M. Chang, and N. M. Hjelm. Rapid clearance of fetal DNA from maternal plasma. *Am. J. Hum. Genet.*, 64(1):218–224, Jan 1999.

[90] J. O. Kitzman, M. W. Snyder, M. Ventura, A. P. Lewis, R. Qiu, L. E. Simmons, H. S. Gammill, C. E. Rubens, D. A. Santillan, J. C. Murray, H. K. Tabor, M. J. Bamshad, E. E. Eichler, and J. Shendure. Noninvasive whole-genome sequencing of a human fetus. *Sci Transl Med*, 4(137):137ra76, Jun 2012.

[91] A. BENDICH, T. WILCZOK, and E. BORENFREUND. CIRCULATING DNA AS A POSSIBLE FACTOR IN ONCOGENESIS. *Science*, 148(3668):374–376, Apr 1965.

[92] R. Catarino, M. M. Ferreira, H. Rodrigues, A. Coelho, A. Nogal, A. Sousa, and R. Medeiros. Quantification of free circulating tumor DNA as a diagnostic marker for breast cancer. *DNA Cell Biol.*, 27(8):415–421, Aug 2008.

[93] R. J. Leary, I. Kinde, F. Diehl, K. Schmidt, C. Clouser, C. Duncan, A. Antipova, C. Lee, K. McKernan, F. M. De La Vega, K. W. Kinzler, B. Vogelstein, L. A. Diaz, and V. E. Velculescu. Development of personalized tumor biomarkers using massively parallel sequencing. *Sci Transl Med*, 2(20):20ra14, Feb 2010.

[94] D. J. McBride, A. K. Orpana, C. Sotiriou, H. Joensuu, P. J. Stephens, L. J. Mudie, E. Hamalainen, L. A. Stebbings, L. C. Andersson, A. M. Flanagan, V. Durbecq, M. Ignatiadis, O. Kallioniemi, C. A. Heckman, K. Alitalo, H. Edgren, P. A. Futreal, M. R. Stratton, and P. J. Campbell. Use of cancer-specific genomic rearrangements to quantify disease burden in plasma from patients with solid tumors. *Genes Chromosomes Cancer*, 49(11):1062–1069, Nov 2010.

[95] S. C. Darby, M. Ewertz, P. McGale, A. M. Bennet, U. Blom-Goldman, D. Bronnum, C. Correa, D. Cutter, G. Gagliardi, B. Gigante, M. B. Jensen, A. Nisbet, R. Peto, K. Rahimi, C. Taylor, and P. Hall. Risk of ischemic heart disease in women after radiotherapy for breast cancer. *N. Engl. J. Med.*, 368(11):987–998, Mar 2013.

[96] S. Misale et al. Emergence of KRAS mutations and acquired resistance to anti-EGFR therapy in colorectal cancer. *Nature*, 486(7404):532–536, Jun 2012.

[97] L. A. Diaz, R. T. Williams, J. Wu, I. Kinde, J. R. Hecht, J. Berlin, B. Allen, I. Bozic, J. G. Reiter, M. A. Nowak, K. W. Kinzler, K. S. Oliner, and B. Vogelstein. The molecular evolution of acquired resistance to targeted EGFR blockade in colorectal cancers. *Nature*, 486(7404):537–540, Jun 2012.

[98] S. J. Dawson, D. W. Tsui, M. Murtaza, H. Biggs, O. M. Rueda, S. F. Chin, M. J. Dunning, D. Gale, T. Forshew, B. Mahler-Araujo, S. Rajan, S. Humphray, J. Becq, D. Halsall, M. Wallis, D. Bentley, C. Caldas, and N. Rosenfeld. Analysis of circulating tumor DNA to monitor metastatic breast cancer. *N. Engl. J. Med.*, 368(13):1199–1209, Mar 2013.

[99] T. Forshew, M. Murtaza, C. Parkinson, D. Gale, D. W. Tsui, F. Kaper, S. J. Dawson, A. M. Piskorz, M. Jimenez-Linan, D. Bentley, J. Hadfield, A. P. May, C. Caldas, J. D. Brenton, and N. Rosenfeld. Noninvasive identification and monitoring of cancer mutations by targeted deep sequencing of plasma DNA. *Sci Transl Med*, 4(136):136ra68, May 2012.

[100] S. N. Cohen, A. C. Chang, and L. Hsu. Nonchromosomal antibiotic resistance in bacteria: genetic transformation of Escherichia coli by R-factor DNA. *Proc. Natl. Acad. Sci. U.S.A.*, 69(8):2110–2114, Aug 1972.

[101] S. N. Cohen, A. C. Chang, H. W. Boyer, and R. B. Helling. Construction of biologically functional bacterial plasmids in vitro. *Proc. Natl. Acad. Sci. U.S.A.*, 70(11):3240–3244, Nov 1973.

[102] K. Kleppe, E. Ohtsuka, R. Kleppe, I. Molineux, and H. G. Khorana. Studies on polynucleotides. XCVI. Repair replications of short synthetic DNA's as catalyzed by DNA polymerases. *J. Mol. Biol.*, 56(2):341–361, Mar 1971.

[103] K. B. Mullis and F. A. Faloona. Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction. *Meth. Enzymol.*, 155:335–350, 1987.

[104] R. K. Saiki, D. H. Gelfand, S. Stoffel, S. J. Scharf, R. Higuchi, G. T. Horn, K. B. Mullis, and H. A. Erlich. Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science*, 239(4839):487–491, Jan 1988.

[105] A. Meyerhans, J. P. Vartanian, and S. Wain-Hobson. DNA recombination during PCR. *Nucleic Acids Res.*, 18(7):1687–1691, Apr 1990.

[106] J. S. Chamberlain, R. A. Gibbs, J. E. Ranier, P. N. Nguyen, and C. T. Caskey. Deletion screening of the Duchenne muscular dystrophy locus via multiplex DNA amplification. *Nucleic Acids Res.*, 16(23):11141–11156, Dec 1988.

[107] A. Edwards, A. Civitello, H. A. Hammond, and C. T. Caskey. DNA typing and genetic mapping with trimeric and tetrameric tandem repeats. *Am. J. Hum. Genet.*, 49(4):746–756, Oct 1991.

[108] A. Edwards, H. A. Hammond, L. Jin, C. T. Caskey, and R. Chakraborty. Genetic variation at five trimeric and tetrameric tandem repeat loci in four human population groups. *Genomics*, 12(2):241–253, Feb 1992.

[109] D. Warne, C. Watkins, P. Bodfish, K. Nyberg, and N. K. Spurr. Tetranucleotide repeat polymorphism at the human beta-actin related pseudogene 2 (ACTBP2) detected using the polymerase chain reaction. *Nucleic Acids Res.*, 19(24):6980, Dec 1991.

[110] L. Albinsson, J. Hedman, and R. Ansell. SKL byter DNA-kit. *Kriminalteknik*, (1), 2011.

[111] M. Nilsson, H. Malmgren, M. Samiotaki, M. Kwiatkowski, B. P. Chowdhary, and U. Landegren. Padlock probes: circularizing oligonucleotides for localized DNA detection. *Science*, 265(5181):2085–2088, Sep 1994.

[112] P. Hardenbol, J. Baner, M. Jain, M. Nilsson, E. A. Namsaraev, G. A. Karlin-Neumann, H. Fakhrai-Rad, M. Ronaghi, T. D. Willis, U. Landegren, and R. W. Davis. Multiplexed genotyping with sequence-tagged molecular inversion probes. *Nat. Biotechnol.*, 21(6):673–678, Jun 2003.

[113] P. Hardenbol et al. Highly multiplexed molecular inversion probe genotyping: over 10,000 targeted SNPs genotyped in a single tube assay. *Genome Res.*, 15(2):269–275, Feb 2005.

[114] Y. Wang, M. Moorhead, G. Karlin-Neumann, N. J. Wang, J. Ireland, S. Lin, C. Chen, L. M. Heiser, K. Chin, L. Esserman, J. W. Gray, P. T. Spellman, and M. Faham. Analysis of molecular inversion probe performance for allele copy number determination. *Genome Biol.*, 8(11):R246, 2007.

[115] G. J. Porreca, K. Zhang, J. B. Li, B. Xie, D. Austin, S. L. Vassallo, E. M. LeProust, B. J. Peck, C. J. Emig, F. Dahl, Y. Gao, G. M. Church, and J. Shendure. Multiplex amplification of large sets of human exons. *Nat. Methods*, 4(11):931–936, Nov 2007.

[116] E. H. Turner, C. Lee, S. B. Ng, D. A. Nickerson, and J. Shendure. Massively parallel exon capture and library-free resequencing across 16 genomes. *Nat. Methods*, 6(5):315–316, May 2009.

[117] J. B. Fan et al. Highly parallel SNP genotyping. *Cold Spring Harb. Symp. Quant. Biol.*, 68:69–78, 2003.

[118] E. Pettersson, M. Lindskog, J. Lundeberg, and A. Ahmadian. Tri-nucleotide threading for parallel amplification of minute amounts of genomic DNA. *Nucleic Acids Res.*, 34(6):e49, 2006.

[119] E. Pettersson, P. Zajac, P. L. Stahl, J. A. Jacobsson, R. Fredriksson, C. Marcus, H. B. Schioth, J. Lundeberg, and A. Ahmadian. Allelotyping by massively parallel pyrosequencing of SNP-carrying trinucleotide threads. *Hum. Mutat.*, 29(2):323–329, Feb 2008.

[120] P. Zajac and A. Ahmadian. Targeted transcript profiling by sequencing. *Sci Rep*, 2:821, 2012.

[121] P. Zajac, C. Oberg, and A. Ahmadian. Analysis of short tandem repeats by parallel DNA threading. *PLoS ONE*, 4(11):e7823, 2009.

[122] F. J. Ghadessy, J. L. Ong, and P. Holliger. Directed evolution of polymerase function by compartmentalized self-replication. *Proc. Natl. Acad. Sci. U.S.A.*, 98(8):4552–4557, Apr 2001.

[123] R. Williams, S. G. Peisajovich, O. J. Miller, S. Magdassi, D. S. Tawfik, and A. D. Griffiths. Amplification of complex gene libraries by emulsion PCR. *Nat. Methods*, 3(7):545–550, Jul 2006.

[124] G. E. Moore. Cramming more components onto integrated circuits. *Electronics*, 38(8), 1965.

[125] K. A. Wetterstrand. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP). *website*, (`www.genome.gov/sequencingcosts`), Accessed Mars 2013.

[126] A. Tiselius. A new apparatus for electrophoretic analysis of colloidal mixtures. *Transactions of the Faraday Society*, 33:524–531, 1937.

[127] R. Markham and J. D. Smith. The structure of ribonucleic acid. I. Cyclic nucleotides produced by ribonuclease and by alkaline hydrolysis. *Biochem. J.*, 52(4):552–557, Dec 1952.

[128] O. Smithies. Zone electrophoresis in starch gels: group variations in the serum proteins of normal human adults. *Biochem. J.*, 61(4):629–641, Dec 1955.

[129] U. E. Loening. The fractionation of high-molecular-weight ribonucleic acid by polyacrylamide-gel electrophoresis. *Biochem. J.*, 102(1):251–257, Jan 1967.

[130] A. Kornberg, I. R. Lehman, M. J Bessman, and E. S Simms. Enzymatic synthesis of desoxyribonucleic acid. *Biochim. Biophys. Acta*, 21:197–198, 1956.

[131] I. R. Lehman. The deoxyribonucleases of Escherichia coli. I. Purification and properties of a phosphodiesterase. *J. Biol. Chem.*, 235:1479–1487, May 1960.

[132] W. Arber. Host-controlled modification of bacteriophage. *Annu. Rev. Microbiol.*, 19:365–378, 1965.

[133] B. Weiss and C. C. Richardson. Enzymatic breakage and joining of deoxyribonucleic acid, I. Repair of single-strand breaks in DNA by an enzyme system from Escherichia coli infected with T4 bacteriophage. *Proc. Natl. Acad. Sci. U.S.A.*, 57(4):1021–1028, Apr 1967.

[134] B. M. Olivera and I. R. Lehman. Linkage of polynucleotides through phosphodiester bonds by an enzyme from Escherichia coli. *Proc. Natl. Acad. Sci. U.S.A.*, 57(5):1426–1433, May 1967.

[135] S. B. Zimmerman, J. W. Little, C. K. Oshinsky, and M. Gellert. Enzymatic joining of DNA strands: a novel reaction of diphosphopyridine nucleotide. *Proc. Natl. Acad. Sci. U.S.A.*, 57(6):1841–1848, Jun 1967.

[136] M. L. Gefter, A. Becker, and J. Hurwitz. The enzymatic repair of DNA. I. Formation of circular lambda-DNA. *Proc. Natl. Acad. Sci. U.S.A.*, 58(1):240–247, Jul 1967.

[137] F. Sanger and A. R. Coulson. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J. Mol. Biol.*, 94(3):441–448, May 1975.

[138] F. Sanger, G. M. Air, B. G. Barrell, N. L. Brown, A. R. Coulson, C. A. Fiddes, C. A. Hutchison, P. M. Slocombe, and M. Smith. Nucleotide sequence of bacteriophage phi X174 DNA. *Nature*, 265(5596):687–695, Feb 1977.

[139] A. M. Maxam and W. Gilbert. A new method for sequencing DNA. *Proc. Natl. Acad. Sci. U.S.A.*, 74(2):560–564, Feb 1977.

[140] C. A. Hutchison. DNA sequencing: bench to bedside and beyond. *Nucleic Acids Res.*, 35(18):6227–6237, 2007.

[141] F. Sanger, S. Nicklen, and A. R. Coulson. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U.S.A.*, 74(12):5463–5467, Dec 1977.

[142] L. M. Smith, J. Z. Sanders, R. J. Kaiser, P. Hughes, C. Dodd, C. R. Connell, C. Heiner, S. B. Kent, and L. E. Hood. Fluorescence detection in automated DNA sequence analysis. *Nature*, 321(6071):674–679, 1986.

[143] P. Nyren. The history of pyrosequencing. *Methods Mol. Biol.*, 373:1–14, 2007.

[144] M. Ronaghi, S. Karamohamed, B. Pettersson, M. Uhlen, and P. Nyren. Real-time DNA sequencing using detection of pyrophosphate release. *Anal. Biochem.*, 242(1):84–89, Nov 1996.

[145] M. Ronaghi, M. Uhlen, and P. Nyren. A sequencing method based on real-time pyrophosphate. *Science*, 281(5375):363, 365, Jul 1998.

[146] F. Mashayekhi and M. Ronaghi. Analysis of read length limiting factors in Pyrosequencing chemistry. *Anal. Biochem.*, 363(2):275–287, Apr 2007.

[147] A. Ahmadian, B. Gharizadeh, A. C. Gustafsson, F. Sterky, P. Nyren, M. Uhlen, and J. Lundeberg. Single-nucleotide polymorphism analysis by pyrosequencing. *Anal. Biochem.*, 280(1):103–110, Apr 2000.

[148] H. Andreasson, A. Asp, A. Alderborn, U. Gyllensten, and M. Allen. Mitochondrial sequence analysis for forensic identification using pyrosequencing technology. *BioTechniques*, 32(1):124–126, Jan 2002.

[149] K. Uhlmann, A. Brinckmann, M. R. Toliat, H. Ritter, and P. Nurnberg. Evaluation of a potential epigenetic biomarker by quantitative methyl-single nucleotide polymorphism analysis. *Electrophoresis*, 23(24):4072–4079, Dec 2002.

[150] M. Margulies et al. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057):376–380, Sep 2005.

[151] Roche Diagnostics Corporation. Products - GS FLX+ System : 454 Life Sciences, a Roche Company. *website*, (`www.454.com/products/gs-flx-system/index.asp`), Accessed Mars 2013.

[152] Roche Diagnostics Corporation. Products - GS Junior System : 454 Life Sciences, a Roche Company. *website*, (`www.454.com/products/gs-junior-system/index.asp`), Accessed Mars 2013.

[153] J. H. Leamon, W. L. Lee, K. R. Tartaro, J. R. Lanza, G. J. Sarkis, A. D. deWinter, J. Berka, M. Weiner, J. M. Rothberg, and K. L. Lohman. A massively parallel PicoTiterPlate based platform for discrete picoliter-scale polymerase chain reactions. *Electrophoresis*, 24(21):3769–3777, Nov 2003.

[154] T. C. Glenn. Field guide to next-generation DNA sequencers. *Mol Ecol Resour*, 11(5):759–769, Sep 2011.

[155] P. C. Ng and E. F. Kirkness. Whole genome sequencing. *Methods Mol. Biol.*, 628:215–226, 2010.

[156] K. E. Wommack, J. Bhavsar, and J. Ravel. Metagenomics: read length matters. *Appl. Environ. Microbiol.*, 74(5):1453–1463, Mar 2008.

[157] Illumina Inc. Solexa Technology. *website*, (`www.illumina.com/technology/solexa-technology.ilmn`), Accessed Mars 2013.

[158] D. R. Bentley, S. Balasubramanian, et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456(7218):53–59, Nov 2008.

[159] B. Toner. In Sequence Survey: Illumina Holds Two-Thirds of Sequencing Market, Splits Desktop Share with Ion PGM. *Genomeweb LLC (website)*, (`www.genomeweb.com/sequencing`), October 2012.

[160] C. Adessi, G. Matton, G. Ayala, G. Turcatti, J. J. Mermod, P. Mayer, and E. Kawashima. Solid phase DNA amplification: characterisation of primer attachment and amplification mechanisms. *Nucleic Acids Res.*, 28(20):E87, Oct 2000.

[161] K. Nakamura, T. Oshima, T. Morimoto, S. Ikeda, H. Yoshikawa, Y. Shiwa, S. Ishikawa, M. C. Linak, A. Hirai, H. Takahashi, M. Altaf-Ul-Amin, N. Ogasawara, and S. Kanaya. Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Res.*, 39(13):e90, Jul 2011.

[162] J. Shendure, G. J. Porreca, N. B. Reppas, X. Lin, J. P. McCutcheon, A. M. Rosenbaum, M. D. Wang, K. Zhang, R. D. Mitra, and G. M. Church. Accurate multiplex polony sequencing of an evolved bacterial genome. *Science*, 309(5741):1728–1732, Sep 2005.

[163] K. J. McKernan et al. Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Res.*, 19(9):1527–1541, Sep 2009.

[164] Life Technologies Corporation. v1.0 SPECIFICATION SHEET: 5500 W SERIES GENETIC ANALYZERS. *downloaded Spec. Sheet*, (`http://www.invitrogen.com/site/us/en/home/brands/Product-Brand/5500-Series-Genetic-Analysis-Systems.html`), Accessed Mars 2013.

[165] J. M. Rothberg et al. An integrated semiconductor device enabling non-optical genome sequencing. *Nature*, 475(7356):348–352, Jul 2011.

[166] B. Merriman, J. M. Rothberg, et al. Progress in ion torrent semiconductor chip based sequencing. *Electrophoresis*, 33(23):3397–3417, Dec 2012.

[167] R. Staden. A strategy of DNA sequencing employing computer programs. *Nucleic Acids Res.*, 6(7):2601–2610, Jun 1979.

[168] Z. Zheng, A. Advani, O. Melefors, S. Glavas, H. Nordstrom, W. Ye, L. Engstrand, and A. F. Andersson. Titration-free 454 sequencing using Y adapters. *Nat Protoc*, 6(9):1367–1376, Sep 2011.

[169] I. Kozarewa, Z. Ning, M. A. Quail, M. J. Sanders, M. Berriman, and D. J. Turner. Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. *Nat. Methods*, 6(4):291–295, Apr 2009.

[170] I. Kozarewa and D. J. Turner. Amplification-free library preparation for paired-end Illumina sequencing. *Methods Mol. Biol.*, 733:257–266, 2011.

[171] M. A. Quail, I. Kozarewa, F. Smith, A. Scally, P. J. Stephens, R. Durbin, H. Swerdlow, and D. J. Turner. A large genome center's improvements to the Illumina sequencing system. *Nat. Methods*, 5(12):1005–1010, Dec 2008.

[172] T. Daley and A. D. Smith. Predicting the molecular complexity of sequencing libraries. *Nat. Methods*, Feb 2013.

[173] A. Adey, H. G. Morrison, X. Xun, J. O. Kitzman, E. H. Turner, B. Stackhouse, A. P. MacKenzie, N. C Caruccio, X. Zhang, and J. Shendure. Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. *Genome biology*, 11(12):R119, 2010.

[174] J. Binladen, M. T. Gilbert, J. P. Bollback, F. Panitz, C. Bendixen, R. Nielsen, and E. Willerslev. The use of coded PCR primers enables high-throughput sequencing of multiple homolog amplification products by 454 parallel sequencing. *PLoS ONE*, 2(2):e197, 2007.

[175] P. Parameswaran, R. Jalili, L. Tao, S. Shokralla, B. Gharizadeh, M. Ronaghi, and A. Z. Fire. A pyrosequencing-tailored nucleotide barcode design unveils opportunities for large-scale sample multiplexing. *Nucleic Acids Res.*, 35(19):e130, 2007.

[176] M. Meyer, U. Stenzel, and M. Hofreiter. Parallel tagged sequencing on the 454 platform. *Nat Protoc*, 3(2):267–278, 2008.

[177] T. J. Albert, M. N. Molla, D. M. Muzny, L. Nazareth, D. Wheeler, X. Song, T. A. Richmond, C. M. Middle, M. J. Rodesch, C. J. Packard, G. M. Weinstock, and R. A. Gibbs. Direct selection of human genomic loci by microarray hybridization. *Nat. Methods*, 4(11):903–905, Nov 2007.

[178] D. T. Okou, K. M. Steinberg, C. Middle, D. J. Cutler, T. J. Albert, and M. E. Zwick. Microarray-based genomic selection for high-throughput resequencing. *Nat. Methods*, 4(11):907–909, Nov 2007.

[179] R. Tewhey, M. Nakano, X. Wang, C. Pabon-Pena, B. Novak, A. Giuffre, E. Lin, S. Happe, D. N. Roberts, E. M. LeProust, E. J. Topol, O. Harismendy, and K. A. Frazer. Enrichment of sequencing targets from the human genome by solution hybridization. *Genome Biol.*, 10(10):R116, 2009.

[180] M. J. Clark, R. Chen, H. Y. Lam, K. J. Karczewski, R. Chen, G. Euskirchen, A. J. Butte, and M. Snyder. Performance comparison of exome DNA sequencing technologies. *Nat. Biotechnol.*, 29(10):908–914, Oct 2011.

[181] Agilent Technologies. HaloPlex - How it Works. *website*, (`http://www.genomics.agilent.com`), Accessed Mars 2013.

[182] J. Eid, S. Turner, et al. Real-time DNA sequencing from single polymerase molecules. *Science*, 323(5910):133–138, Jan 2009.

[183] Pacific Biosciences. PacBio RS - Single Molecule Real Time Sequencing. *downloaded brochure*, (`http://www.pacificbiosciences.com/brochure`), Accessed Mars 2013.

[184] EC Hayden. Nanopore genome sequencer makes its debut. *Nature*, 10, 2012.

[185] Y. Erlich, K. Chang, A. Gordon, R. Ronen, O. Navon, M. Rooks, and G. J. Hannon. DNA Sudoku–harnessing high-throughput sequencing for multiplexed specimen analysis. *Genome Res.*, 19(7):1243–1253, Jul 2009.

[186] M. Galan, E. Guivier, G. Caraux, N. Charbonnel, and J. F. Cosson. A 454 multiplex sequencing method for rapid and reliable genotyping of highly polymorphic genes in large-scale studies. *BMC Genomics*, 11:296, 2010.

[187] P. Savolainen, Y. P. Zhang, J. Luo, J. Lundeberg, and T. Leitner. Genetic evidence for an East Asian origin of domestic dogs. *Science*, 298(5598):1610–1613, Nov 2002.

[188] J. F. Pang, C. Kluetsch, X. J. Zou, A. B. Zhang, L. Y. Luo, H. Angleby, A. Ardalan, C. Ekstrom, A. Skollermo, J. Lundeberg, S. Matsumura, T. Leitner, Y. P. Zhang, and P. Savolainen. mtDNA data indicate a single origin

for dogs south of Yangtze River, less than 16,300 years ago, from numerous wolves. *Mol. Biol. Evol.*, 26(12):2849–2864, Dec 2009.

[189] S. Lundin, H. Stranneheim, E. Pettersson, D. Klevebring, and J. Lundeberg. Increased throughput by parallelization of library preparation for massive sequencing. *PLoS ONE*, 5(4):e10029, 2010.

[190] O. Harismendy, P. C. Ng, R. L. Strausberg, X. Wang, T. B. Stockwell, K. Y. Beeson, N. J. Schork, S. S. Murray, E. J. Topol, S. Levy, and K. A. Frazer. Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biol.*, 10(3):R32, 2009.

[191] J. Sandberg, B. Werne, M. Dessing, and J. Lundeberg. Rapid flow-sorting to simultaneously resolve multiplex massively parallel sequencing products. *Sci Rep*, 1:108, 2011.

[192] F. R. Schumacher et al. Genome-wide association study identifies new prostate cancer susceptibility loci. *Hum. Mol. Genet.*, 20(19):3867–3875, Oct 2011.

[193] F. C. Hsu, J. Sun, F. Wiklund, S. D. Isaacs, K. E. Wiley, L. D. Purcell, Z. Gao, P. Stattin, Y. Zhu, S. T. Kim, Z. Zhang, W. Liu, B. L. Chang, P. C. Walsh, D. Duggan, J. D. Carpten, W. B. Isaacs, H. Gronberg, J. Xu, and S. L. Zheng. A novel prostate cancer susceptibility locus at 19q13. *Cancer Res.*, 69(7):2720–2723, Apr 2009.

[194] A. Chien, D. B. Edgar, and J. M. Trela. Deoxyribonucleic acid polymerase from the extreme thermophile Thermus aquaticus. *J. Bacteriol.*, 127(3):1550–1557, Sep 1976.

[195] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, Aug 2009.

[196] A. Wysoker, K. Tibbetts, and T. Fennell. Picard. *website*, (`http://picard.sourceforge.net`), Accessed Mars 2013.

[197] A. Narayan, N. J. Carriero, S. N. Gettinger, J. Kluytenaar, K. R. Kozak, T. I. Yock, N. E. Muscato, P. Ugarelli, R. H. Decker, and A. A. Patel. Ultra-sensitive measurement of hotspot mutations in tumor DNA in blood using error-suppressed multiplexed deep sequencing. *Cancer Res.*, 72(14):3492–3498, Jul 2012.

[198] D. Dressman, H. Yan, G. Traverso, K. W. Kinzler, and B. Vogelstein. Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. *Proc. Natl. Acad. Sci. U.S.A.*, 100(15):8817–8822, Jul 2003.

[199] M. Murtaza, S. J. Dawson, D. W. Tsui, D. Gale, T. Forshew, A. M. Piskorz, C. Parkinson, S. F. Chin, Z. Kingsbury, A. S. Wong, F. Marass, S. Humphray, J. Hadfield, D. Bentley, T. M. Chin, J. D. Brenton, C. Caldas, and N. Rosenfeld. Non-invasive analysis of acquired resistance to cancer therapy by sequencing of plasma DNA. *Nature*, Apr 2013.

[200] P. M. Holland, R. D. Abramson, R. Watson, and D. H. Gelfand. Detection of specific polymerase chain reaction product by utilizing the 5'->3' exonuclease activity of Thermus aquaticus DNA polymerase. *Proc. Natl. Acad. Sci. U.S.A.*, 88(16):7276–7280, Aug 1991.

[201] K. J. Livak. Allelic discrimination using fluorogenic probes and the 5' nuclease assay. *Genet. Anal.*, 14(5-6):143–149, Feb 1999.

[202] E. Heitzer, P. Ulz, J. Belic, S. Gutschi, F. Quehenberger, K. Fischereder, T. Benezeder, M. Auer, C. Pischler, S. Mannweiler, M. Pichler, F. Eisner, M. Haeusler, S. Riethdorf, K. Pantel, H. Samonigg, G. Hoefler, H. Augustin, J. B. Geigl, and M. R. Speicher. Tumor associated copy number changes in the circulation of patients with prostate cancer identified through whole-genome sequencing. *Genome Med*, 5(4):30, Apr 2013.

[203] E. Heitzer, M. Auer, E. M. Hoffmann, M. Pichler, C. Gasch, P. Ulz, S. Lax, J. Waldispuehl-Geigl, O. Mauermann, S. Mohan, G. Pristauz, C. Lackner, G. Hofler, F. Eisner, E. Petru, H. Sill, H. Samonigg, K. Pantel, S. Riethdorf, T. Bauernhofer, J. B. Geigl, and M. R. Speicher. Establishment of tumor-specific copy number alterations from plasma DNA of patients with cancer. *Int. J. Cancer*, Jan 2013.